Tânia Rafaela Gonçalves da Rocha

Risco de Sépsis/Meningite em Prematuros de muito baixo peso

Risk of Sepsis/Meningitis in very low birth weight Newborns

Tânia Rafaela Gonçalves da Rocha

Risco de Sépsis/Meningite em Prematuros de muito baixo peso

Risk of Sepsis/Meningitis in very low birth weight Newborns

"*Morre de ter ousado na água amar o fogo.*"

— Eugénio de Andrade

Universidade de Aveiro Departamento de Matemática
2017

Tânia Rafaela Gonçalves da Rocha

Risco de Sépsis/Meningite em Prematuros de muito baixo peso

Risk of Sepsis/Meningitis in very low birth weight Newborns

Relatório de estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Doutora Isabel Maria Simões Pereira, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Dedico este trabalho a todos aqueles que me ajudaram a tornar na mulher que sou hoje.

Júri

Presidente

Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

Vogais

Doutor Bruno Miguel Alves Fernandes do Gago
Professor Auxiliar Convidado do Departamento de Ciências Médicas da Universidade de Aveiro

Doutora Isabel Maria Simões Pereira
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

Agradecimentos

Agradeço aos Professores Isabel Pereira, Adelaide Freitas, Pedro Macedo e Luís Silva por todo o apoio prestado. Agradeço, também, a toda a equipa da empresa Mhii, onde estagiei, pela forma como fui recebida, acolhida e estimulada dia após dia para desenvolver o trabalho que aqui descrevo.

Palavras Chave

ANOVA, colinearidade, curva ROC, imputação, previsão, regressão logística, seleção de variáveis, sépsis, valor omisso.

Resumo

Vivemos, atualmente, num mundo repleto de números e dados, que, sendo devidamente recolhidos, organizados e interpretados podem sustentar importantes decisões nos mais variados setores.

Foi também com este propósito que este trabalho foi desenvolvido, de forma a poder auxiliar a comunidade médica a travar a principal causa de morbilidade e mortalidade neonatal - Sépsis. Assim, será descrito ao longo de todo este trabalho, um modelo com um boa capacidade preditiva do risco de Sépsis/Meningite em Prematuros de muito baixo peso.

A base de dados cuja análise está na base deste trabalho foi cedida pela Sociedade Portuguesa de Neonatologia e é composta pelos registos de 7506 indivíduos relativamente a 128 variáveis.

Foram descartadas cerca de metade das variáveis (relacionadas com datas e registos associados a Transferência e Internamento) uma vez que não teriam qualquer impacto nos modelos que seriam construídos. Foi aplicado o método de imputação dos k-vizinhos mais próximos para o tratamento dos valores omissos e o método *Stepwise* para a escolha das variáveis a considerar nos modelos. Conseguiu-se reduzir a colinearidade dos dados pela análise dos fatores de inflação da variância (VIF) e melhorar, assim, a capacidade preditiva do modelo.

Recorreu-se ao modelo de regressão logística uma vez que a variável resposta - Sépsis/Meningite é uma variável dicotómica e analisou-se, também, modelos de regressão por penalização - *Lasso*, *Ridge* e *Elastic Net*, de forma a tentar melhorar a capacidade preditiva dos modelos.

Keywords ANOVA, collinearity, imputation, logistic regression, missing value, prediction, ROC curve, selection of variables, Sepsis.

Abstract

We are living, today, in a world full of numbers and data, which, when properly collected, organized and interpreted, can sustain important decisions in the most varied sectors. It was also for this purpose that this work was developed, in order to help the medical community to stop the main cause of neonatal morbidity and mortality - Sepsis. Thus, a model with a good predictive capacity of the risk of Sepsis/Meningitis in very low birth weight infants will be described throughout this study.

The database whose analysis is the basis of this work was provided by the Portuguese Society of Neonatology and is composed by the records of 7506 individuals regarding 128 variables.

About half of the variables (related to dates and records associated with relocation) were discarded because they would have no impact on the models that would be constructed. The imputation method of the nearest k-neighbors was applied for the treatment of missing values and the Stepwise method for choosing the variables to consider in the models. It was possible to reduce the collinearity of the data by analyzing the variance inflation factors (VIF) and thus improve the predictive capacity of the model. The logistic regression model was used because the response - Sepsis/Meningitis variable is a dichotomous variable, and regression models were also analyzed by penalization - Lasso, Ridge and Elastic Net, in order to improve the predictive capacity of the models.

Conteúdo

Lista de Tabelas

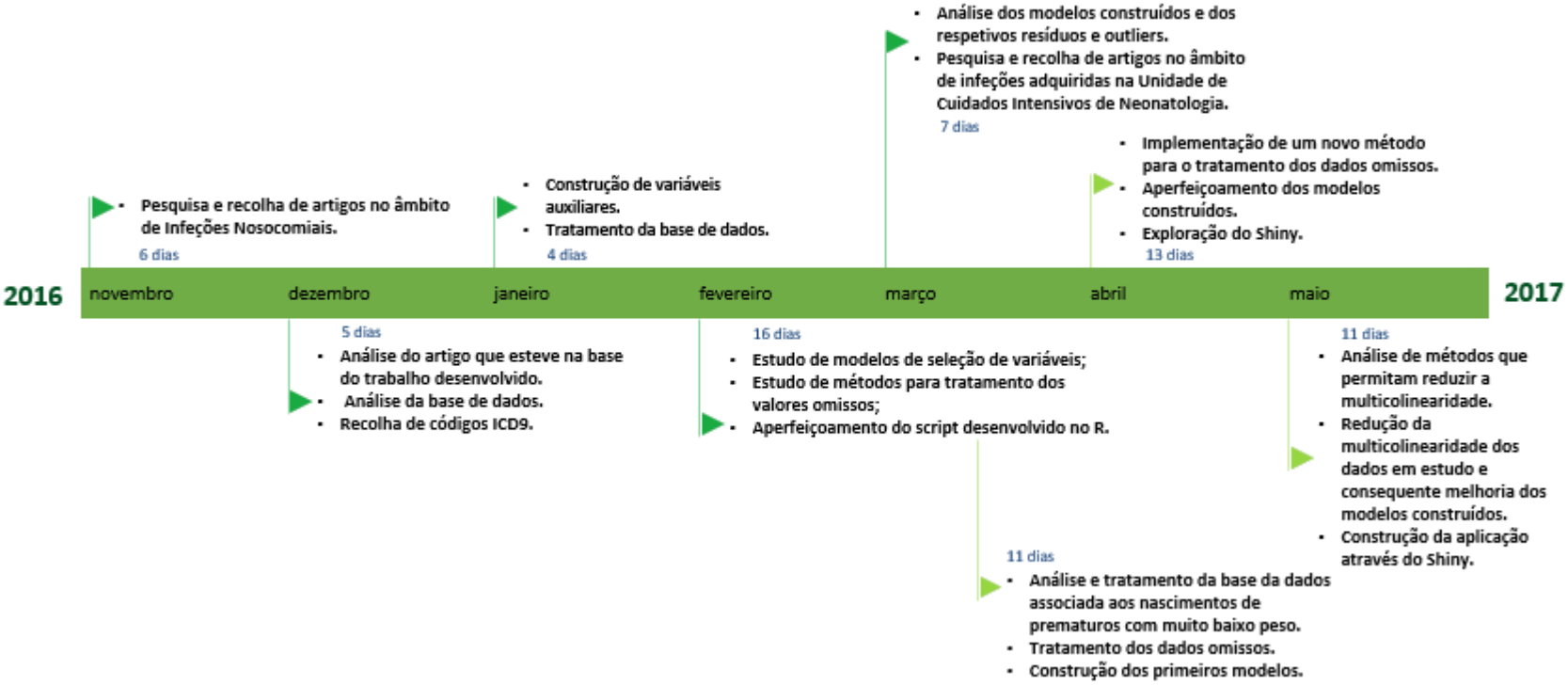
Capítulo 1

Cronograma

Segue-se uma descrição sumária do trabalho desenvolvido ao longo do estágio curricular do mestrado em Matemática e Aplicações, ramo de especialização Estatística e Otimização na empresa Mhii com datas de início a 10/11/2016 e término a 31/05/2017.
Analisando o cronograma - Figura 1.1 verifica-se que mais de 50% do tempo dedicado ao estágio curricular foi aplicado no tratamento da primeira base de dados cedida pela Mhii, com o objetivo de estudar o risco de adquirir uma infeção nosocomial (infeção adquirida em ambiente hospital). Grande parte do trabalho foi desenvolvido com sucesso mas não foi possível concluí-lo dada a impossibilidade de encontrar as centenas de códigos ICD9 (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde - *International Statistical Classification of Diseases and Related Health Problems* que fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças) em falta junto de um médico codificador.

O trabalho desenvolvido durante o tratamento da primeira base de dados foi extremamente útil no tratamento da segunda base de dados cedida pela Mhii, a qual reune os registos associados a recém-nascidos de muito baixo peso em Portugal, entre 01.01.2010 e 31.01.2017. O trabalho desenvolvido nesta segunda fase teve como objetivo a construção de um modelo com um boa capacidade preditiva do risco de Sépsis/Meningite em Prematuros de muito baixo peso, o qual será descrito ao longo deste trabalho.

Figura 1.1: Cronograma do trabalho desenvolvido



Como consequência do trabalho desenvolvido elaborou-se este relatório de estágio que está organizado da seguinte forma:

• no Capítulo 2 faz-se uma breve apresentação da empresa onde decorreu o estágio;

• no Capítulo 3 faz-se uma breve descrição da infeção, Sépsis, que foi alvo de estudo neste trabalho;

• nos Capítulos 4 e 5 faz-se uma apresentação dos conceitos matemáticos que foram preponderantes na construção dos diversos modelos que serão descritos posteriormente;

• no Capítulo 6 faz-se uma breve descrição dos métodos utilizados no tratamento dos valores omissos da base de dados que está na base deste trabalho;

• no Capítulo 7 faz-se uma análise da base de dados que foi alvo de estudo;

• no Capítulo 8 faz-se uma apresentação dos melhores modelos que foram construídos nas 5 etapas que serão apresentadas, assim como a análise dos erros e outliers do último modelo construído;

• no Capítulo 9 faz-se uma apresentação e descrição da aplicação web que foi construída com base no melhor modelo;

• no Capítulo 10 enumeram-se as conclusões retiradas deste trabalho assim como alguns desafios futuros.

Capítulo 2

Mhii

Giving health its value

Este trabalho foi desenvolvido segundo a orientação da equipa que compõe a empresa Mhii filiada em São João da Madeira e à qual muito agradeço pela forma como fui recebida e acolhida. A Mhii é uma empresa de tecnologia direcionada para a inovação na Saúde em instituições de saúde públicas e privadas, como Hospitais, Clínicas, Gabinetes, ONGs e associações sectoriais. Foi fundada em setembro de 2015 e tem equipas multidisciplinares focadas em *giving health its value* e com experiência em:

• Engenharia de Dados: construção, integração e manutenção de ferramentas e plataformas para o processamento de dados em tempo real;

• Inteligência de negócios: recolha, organização, análise e monitorização de dados que permitem um maior suporte a gestão de negócios;

• Analítica Avançada e Design de Interação: uso de *data mining* robusto, *machine learning*, otimização e técnicas de simulação de forma a encontrar padrões ocultos e detectar anomalias nos dados, prever eventos, avaliar potenciais riscos e melhorar recursos já existentes com abordagens de concepção centradas na experiência do utilizador.

A Mhii está sediada em Portugal e, recentemente, fundiu-se com a Prológica, uma das mais antigas empresas portuguesas que operam na área das Tecnologias da Informação e Comunicação (TIC) desde 1984.

Capítulo 3

Sépsis

A Sépsis neonatal é uma importante causa de morbilidade e mortalidade neonatal.
Dos 4 milhões de mortes neonatais estimadas por ano em todo o mundo, mais de um terço são causadas por infecções graves, sendo que um quarto é devida a Sépsis neonatal/Pneumonia.
Se diagnosticada e tratada precocemente, torna-se possível salvar a maioria dos recém-nascidos com Sépsis neonatal (Mirco [9]).

A Sépsis neonatal é designada como precoce, se o início do aparecimento dos sintomas ocorre nos primeiros três dias de vida, ou então como tardia, se for entre o quarto e centésimo vigésimo dias de vida (Hornik et al.[6]).
Cerca de 85% dos recém-nascidos diagnosticados com Sépsis neonatal precoce apresentam sintomas nas primeiras 24h de vida, 5% no segundo dia de vida e uma pequena percentagem até às 72h de vida (Mirco[9]).
A Sépsis precoce pode ser fulminante, com uma taxa de mortalidade que varia entre 3-50%, está associada ao aparecimento simultâneo de Pneumonia e relacionada com fatores gestacionais e/ou periparto (Mirco[9]).
Os microrganismos que se encontram mais frequentemente associados com a Sépsis precoce são: o *Streptococcus* do grupo B, *Escherichia coli*, *Staphylococcus* coagulase-negativo, *Haemophillus influenzae* e *Listeria monocytogenes*.

Por sua vez, a Sépsis tardia possui uma taxa de mortalidade mais baixa do que a da

precoce, variando entre 2-40%, está associada ao aparecimento simultâneo de Meningite e relacionada com os cuidados de saúde (Mirco[9]).

Os microrganismos que se encontram mais frequentemente associados com a Sépsis tardia e Meningite incluem os que já foram referidos anteriormente para a Sépsis precoce e outros como o *Staphylococcus aureus* e *Pseudomonas aeruginosa*.

Seguem-se alguns factores de risco da Sépsis (de uma forma generalizada) (Mirco[9]):

- índice de Apgar <6 aos 5 minutos (que sugere sofrimento do recém-nascido antes ou na altura do parto);

O sistema de pontuação Apgar foi criado em 1952 por Virginia Apgar, médica e anestesista, com o objetivo de avaliar a condição do recém-nascido aos 1 e 5 minutos de vida. Os recém-nascidos são avaliados com base em cinco variáveis: frequência cardíaca, esforço respiratório, tónus (tensão) muscular e irritabilidade reflexa. A cada variável é atribuído entre 0 a 2 pontos, como indica a tabela 3.1 e do somatório das pontuações resulta o índice de Apgar (atingindo assim um valor máximo de 10). Note-se que quanto maior for o índice Apgar melhor será o estado de saúde do recém-nascido (Montgomery[10]).

Tabela 3.1: Tabela para cálculo do índice Apgar

Pontos	0	1	2
Frequência cardíaca	Ausente	<100/minuto	>100/minuto
Respiração	Ausente	Fraca, irregular	Forte/Choro
Tónus Muscular	Flácido	Flexão de pernas e braços	Movimento ativo/Boa flexão
Cor	Cianótico/Pálido	Cianose de extremidades	Rosado
Irritabilidade reflexa	Ausente	Algum movimento	Espirros/Algum choro

- parto prematuro (menos de 35 semanas);
- infecção do trato urinário;
- febre materna intraparto;
- sinais de sofrimento fetal (taquicardia >160 bat/min, líquido amniótico fétido);
- recém-nascidos com baixo peso, pré-termo e pequenos para a idade gestacional;
- síndrome de dificuldade respiratória;
- aumento da necessidade de oxigénio;

• episódios de bradicardia (frequência cardíaca <80 bat/min) ou de taquicardia (frequência cardíaca >200 bat/min);

• alterações metabólicas;

• hemorragia intracraniana;

• parto traumático;

• entre outros.

3.1 Prevenção

A intervenção mais simples que previne ou diminui as infeções neonatais é a correta lavagem/desinfeção das mãos.

Os cuidados pré-natais prestados às grávidas ajudam a reduzir os taxas de recém-nascidos prematuros e consequentemente o risco de vir a ser diagnosticado Sépsis.

O uso de corticóides pré-natais em grávidas com ameaça de parto prematuro e uso de surfactante exógeno nessas crianças diminui substancialmente o aparecimento de síndrome de dificuldade respiratória nos recém-nascidos e consequentemente o risco de vir a ser diagnosticado Sépsis (Mirco[9]).

Foi também com este propósito, de minimizar o risco de vir a ser diagnostico Sépsis, que este trabalho foi desenvolvido, mais focado em prematuros de muito baixo peso. Numa fase inicial o estudo iria incidir separadamente em ambas as Sépsis, precoce e tardia, mas optou-se, numa fase posterior, que este fosse generalizado apenas em Sépsis. Note-se que a base de dados que está na origem de todo este trabalho tinha inicialmente as duas variáveis Sépsis precoce ou tardia associadas a Meningite, o que poderá estar a causar algum tipo de ruído nos resultados finais obtidos (uma vez que como foi referido anteriormente a Sépsis precoce está associada ao aparecimento simultâneo de Pneumonia enquanto que a Sépsis tardia com Meningite). Os vários modelos que foram construídos e que aqui serão descritos darão, em tempo real, a probabilidade que o prematuro terá a vir a adquirir Sépsis. Serão, assim, boas ferramentas preditivas e permitirão, com uma margem de erro, que o diagóstico possa ser feito mais precocemente e de uma forma mais acertiva.

Capítulo 4

Regressão Logística

Um dos problemas mais interessantes em Estatística é descobrir e avaliar a relação entre duas ou mais variáveis. Caso essa relação exista, objetivo é pois descrevê-la formalmente através de uma equação matemática.
Na realidade, o conhecimento da equação matemática que relaciona as variáveis é de enorme importância, não só pela informação científica que intrinsecamente contém, mas também pela importância prática de permitir a previsão dos valores de uma variável dependente com base nos valores de uma ou mais variáveis independentes (Cordeiro e Magalhães[1]).

4.1 Introdução

O termo "regressão" foi proposto pela primeira vez por Francis Galton em 1885 num estudo onde demonstrou que a altura dos filhos não tende a refletir a altura dos pais mas sim a regredir para a média da população (Marôco[8]).
Atualmente, o termo "regressão" define um grande conjunto de técnicas no campo da Estatística que permite modelar relações entre variáveis e obter previsões da variável resposta/dependente a partir das variáveis independentes/preditoras (Marôco[8]).
A regressão linear, tipo de regressão usada mais frequentemente, não será utilizada no âmbito deste trabalho e distingue-se da regressão logística pelo facto da variável reposta ser tomada como contínua e não dicotómica ou polidicotómica (o que carateriza a regressão logística).
Todavia, as técnicas usadas em ambos os métodos são muito semelhantes (Hosmer et al.[7]).

4.2 Regressão logística múltipla

A regressao logística é designada como simples quando existe apenas uma variável independente e múltipla caso exista pelo menos duas. Assim, será dado mais ênfase à regressão logística múltipla pois será esse o método que será aplicado à base da dados que está na origem deste trabalho.

Genericamente, um modelo de regressão logística múltiplo com p variáveis independentes, $x = (x_1, \ldots, x_p)$, é expresso por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}.$$

Linearizando:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$g(x) = \ln\left[\frac{\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}}{1 - \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}}\right] = \ln\left[\frac{\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}}{\frac{1}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}}\right] = \ln\left[e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}\right].$$

Assim,

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Esta transformação também é designada por transformação *logit* da regressão logística múltipla e tem muitas das propriedades desejáveis dos modelos de regressão linear, apresenta linearidade nos parâmetros e deve ser contínua em $\mathbb{R}$ (dependendo dos valores de x).

Note-se que, enquanto que nos modelos de regressão linear a variável resposta pode assumir qualquer valor, nos modelos de regressão logística a variável resposta é dicotómica (habitualmente a presença da característica definida pela variável em questão é identificada com o número 1 e, por sua vez, a sua ausência é identificada com o número 0).

4.2.1 Ajustamento do modelo de regressão logística

Para ajustar o modelo descrito anteriormente pela equação

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

é necessário estimar os parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ pelo método de máxima verosimilhança.

Tratando-se de um modelo de regressão logística, a variável resposta toma apenas um de dois

valores, 0 ou 1, ou seja, cada observação Y_i é encarada como um ensaio de Bernoulli tal que $Y_i \sim B(1, \pi)$.

Assim,

$$P(Y = y_i) = \pi^{y_i}(1-\pi)^{1-y_i}.$$

Supondo que todas as observações são independentes, então a função de verosimilhança é dada por

$$l(\beta) = P(Y = y_1, \ldots, Y = y_n) = P(Y = y_1) \times \ldots \times (Y = y_n) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}.$$

O máximo desta função verifica-se quando $\partial l/\partial\beta = 0$ e $\partial^2 l/\partial^2\beta < 0$.

Atendendo que é mais fácil derivar uma soma do que um produto, tome-se

$$ln(l(\beta)) = \sum_{i=1}^{n} \left[y_i ln(\pi_i) - (1-y_i) ln(1-\pi_i) \right].$$

Assim,

$$ln(l(\beta)) = \sum_{i=1}^{n} \left[y_i(\beta_0 + y_i\beta_1 x_{i1} + \ldots + \beta_p x_{ip}) - ln(1 + e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}) \right].$$

O máximo de $ln(l(\beta))$ ocorre com o vetor dos coeficientes para o qual $\partial ln(l)/\partial\beta = 0$. Como o $ln(l)$ é uma transformação monótona em l, o vetor que maximiza $ln(l)$ é também o vetor que maximiza l. Porém, este sistema de $p+1$ equações não tem uma solução analítica, pelo que β é estimado iterativamente por um algoritmo computacional que maximiza o $ln(l)$.

São também calculadas as estatísticas que servem para avaliar a qualidade do modelo e a significância dos coeficientes de regressão (Marôco[8]).

Note-se que as $p+1$ equações da função log verosimilhança serão expressas da seguinte forma:

$$\sum_{i=1}^{n} [y_i - \pi(x_i] = 0 \text{ e } \sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i] = 0$$

para $j = 1, 2, \ldots, p$.

Para além das estimativas dos parâmetros, podem ser calculadas as variâncias e covariâncias do erro associado a cada uma das estimativas, a partir das segundas derivadas da função log

verosimilhança.

Assim,

$$\partial^2 ln(l)/\partial\beta_j^2 = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1-\pi_i)$$

e

$$\partial^2 ln(l)/\partial\beta_j\partial\beta_l = -\sum_{i=1}^{n} x_{ij}x_{il}\pi_i(1-\pi_i)$$

onde $j, l = 0, \ldots, p$ e $\pi_i = \pi(x_i)$.

Seja $I(\beta)$ uma matriz quadrada de dimensão $p+1$ constituída pelos valores das duas equações anteriores e designada por Matriz de Informação Observada. As variâncias e covariâncias dos coeficientes estimados são obtidos a partir da matriz inversa de $I(\beta)$, ou seja, $Var(\beta) = I^{-1}(\beta)$. Assim, o j-ésimo elemento da diagonal da matriz $I^{-1}(\beta)$ representará $Var(\beta_j)$ (variância de $\widehat{\beta_j}$) e um elemento arbitrário fora da diagonal representará à covariância entre esses coeficientes estimados (covariância entre $\widehat{\beta_j}$ e $\widehat{\beta_l}$). Os estimadores da variância e covariância, identificados respetivamente por $\widehat{Var}(\widehat{\beta_j})$ e $\widehat{Cov}(\widehat{\beta_i}, \widehat{\beta_j})$, para $j, l = 0, \ldots, p$, são obtidos a partir da matriz $I^{-1}(\beta)$, substituindo o parâmetro β pelo seu estimador $\widehat{\beta}$.

Note-se que o desvio padrão estimado associado às estimativas dos parâmetros é definido da seguinte forma:

$$\widehat{SE}(\widehat{\beta_j}) = [\widehat{Var}(\widehat{\beta_j})]^{1/2}.$$

4.2.2 Critério de informação do modelo

Um dos critérios mais utilizados para comparar a qualidade de dois modelos é o critério de *Akaike*, *Akaike Information Criterion* - AIC, definido como:

$$AIC = -2LL(\beta, \theta) + 2p$$

onde LL representa o logaritmo das funções de verosimilhança e p é o número de variáveis do modelo com poder preditivo.

O AIC penaliza os modelos com maior número de parâmetros e o melhor modelo é aquele que apresenta menor AIC (Marôco[8]).

4.2.3 Teste à significância dos coeficientes do modelo de regressão logística

De forma a verificar-se se existe pelo menos uma variável independente linearmente relacionada com o $logit(\pi_j)$ recorre-se ao Teste de Wald. Neste teste pretende-se testar se um determinado coeficiente é ou não nulo, mediante os valores estimados dos restantes coeficientes. Assim, realiza-se o teste de hipóteses que se segue:

$$H_0 : \beta_i = 0 | \beta_0, ..., \beta_p$$

vs

$$H_1 : \beta_i \neq 0 | \beta_0, ..., \beta_p$$

para $i = 1, ..., p$.

A estatística de teste é

$$T_{Wald_i} = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}$$

onde $\hat{\beta}_i$ é o estimador de β_i e $\hat{SE}(\hat{\beta}_i)$ é o estimador do erro-padrão de β_i.

Esta estatística aproxima-se assintoticamente da distribuição N(0,1) quando a distribuição da amostra é relativamente grande e a hipótese H_0 é rejeitada para cada um dos testes aos β_i quando o *p-value* é inferior ou igual ao nível de significância fixado (é habitualmente 0.05).

4.2.4 Teste de ajustamento do modelo de regressão logística

O modelo de regressão logística depois de devidamente ajustado deve ser avaliado quanto à sua significância recorrendo, para isso, ao seguinte teste de hipóteses (Hosmer et al.[7]):

$$H_0 : \text{O modelo ajusta-se aos dados}$$

vs

$$H_1 : \text{O modelo não se ajusta aos dados}$$

A estatística de teste clássica é o Qui-quadrado de Pearson:

$$X_P^2 = \Sigma_{j=1}^{J} \left(\frac{y_j - n_j \hat{\pi}_j}{\sqrt{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \right)^2 = \Sigma_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$$

onde J representa o número de células em que os dados se encontram agrupados, y_j é o

número de sucessos na célula j, $\hat{\pi}_j$ é a probabilidade estimada da célula j, n_j é o número de elementos da célula j e O_j e E_j são o número de sucessos observados e esperados na célula j, respetivamente.

Assim, para amostras grandes, X_p^2 apresenta distribuição assintótica Qui-quadrado com $J - p - 1$ graus de liberdade e rejeita-se H_0 se o p-*value* foi inferior ou igual ao nível de significância fixado (é habitualmente 0.05).

Note-se que para que a estatística de teste seja válida o número de células não deve ser aproximadamente igual ao número de elementos da amostra.

4.2.5 Seleção de varáveis com poder preditivo

Os três métodos utilizados para a seleção das variáveis foram os seguintes (para além dos modelos construídos através análise da importância das mesmas):

- Seleção *Forward*: o modelo é construído inicialmente com apenas uma variável e é adicionado em cada iteração mais uma variável; as variáveis que apresentem maior correlação com a variável dependente são sucessivamente incluídas;

- Seleção *Backward*: o pressuposto deste método é contrário ao anterior, ou seja, é inicializado com todas as variáveis e são removidas uma a uma até obter um modelo simples; as variáveis que apresentem menor correlação com a variável dependente são sucessivamente excluídas;

- Seleção *Stepwise*: combinação dos dois métodos anteriores.

Optou-se por utilizar apenas o método *Stepwise* ao longo do trabalho que será posteriormente descrito uma vez que é uma combinação dos dois métodos anteriores.

4.2.6 Análise da variância

A avaliação do modelo com base na tabela ANOVA (o nome vem de *Analysis of Variance*) com pelo menos dois fatores (pelo menos duas variáveis independentes) é a que será alvo de estudo ao longo deste trabalho. Esta análise é bastante complexa pois pretende-se estudar o efeito não só de cada um dos fatores mas e também a possível influência que cada um dos fatores pode exercer sobre a resposta da variável dependente ao outro fator (Marôco[8]). Este efeito é designado por *efeito de interação ou moderação* entre fatores.

Será considerado, por uma questão de simplificação, a existência de apenas dois fatores mas a ANOVA de 3 ou mais fatores é interpretada como uma extensão da ANOVA a dois fatores. Considerem-se, então, os fatores A com a níveis $i = 1, ..., a$ e B com b níveis $j = 1, ..., b$ e, para simplificação da exposição, que cada uma das combinações dos níveis (isto é cada um dos tratamentos ou amostras) de ambos os fatores possuam r repetições (mesma dimensão). Assim, o modelo teórico da ANOVA é descrito como

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr},$$

em que α_i é o efeito do fator A, β_j é o efeito do fator B, γ_{ij} representa a interação entre ambos os fatores, μ representa a média global e ε_{ijr} representa os erros [toma-se $\varepsilon_{ijr} \sim N(0, \sigma)$]. Desta forma, o modelo ANOVA descrito a partir das observações amostrais é representado da seguinte forma (através das estimativas de cada uma das parcelas da expressão anterior):

$$y_{ijr} = \overline{y} + (\overline{y}_i - \overline{y}) + (\overline{y}_j - \overline{y}) + (\overline{y}_{ij} - \overline{y}_i - \overline{y}_j + \overline{y}) + (y_{ijr} - \overline{y}_{ij})$$

em que $\overline{y}$ representa a média geral da amostra, $\overline{y}_i$ representa a média de cada nível da variável em questão e $\overline{y}_{ij}$ representa a média por amostra.

O objetivo desta análise é testar se as médias para cada nível dos fatores A e B (ou os demais fatores) são ou não iguais, ou seja, testar o seguinte conjunto de testes de hipóteses:

1. $H_0^A : \mu_1 = \mu_2 = ... = \mu_a$ vs $H_1^A : \exists i, j : \mu_i \neq \mu_j (i \neq j; i, j = 1, ..., a)$
2. $H_0^B : \mu_1 = \mu_2 = ... = \mu_b$ vs $H_1^B : \exists i, j : \mu_i \neq \mu_j (i \neq j; i, j = 1, ..., b)$
3. $H_0^\gamma : \gamma_{ij} = 0$ (não existe interação entre os fatores) vs $H_1^\gamma : \gamma_{ij} \neq 0$ (i=1,...,a; j=1,...,b) (existe interação entre os fatores)

Veja-se a tabela 4.1, tabela resumo da ANOVA a dois fatores.

Fonte de variação	Soma de Quadrados	g.l.	Quadrados médios	F
Fator A	$SQF_A = b \times r \times \Sigma_{i=1}^{a}(\overline{Y}_i - \overline{Y})^2$	a-1	$QMF_A = \frac{SQF_A}{a-1}$	$F_A = \frac{QMF_A}{QME}$ $F_A \sim F_{a-1,(r-1)ab}$
Fator B	$SQF_B = a \times r \times \Sigma_{j=1}^{b}(\overline{Y}_j - \overline{Y})^2$	b-1	$QMF_B = \frac{SQF_B}{b-1}$	$F_B = \frac{QMF_B}{QME}$ $F_B \sim F_{b-1,(r-1)ab}$
Interação	$SQ_{AB} = r \times \Sigma_{i=1}^{a}\Sigma_{j=1}^{b}(\overline{Y}_{ij} - \overline{Y}_i - \overline{Y}_j + \overline{Y})^2$	(a-1)(b-1)	$QMF_{A\times B} = \frac{SQ_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{QMF_{A\times B}}{QME}$ $F_{AB} \sim F_{(a-1)(b-1),(r-1)ab}$
Erro	$SQE = \Sigma_{i=1}^{a}\Sigma_{j=1}^{b}\Sigma_{l=1}^{r}(Y_{ijl} - \overline{Y}_{ij})^2$ (r-1)ab	$\frac{SQE}{(r-1)ab}$		
Total	$SQT = \Sigma_{i=1}^{a}\Sigma_{j=1}^{b}\Sigma_{l=1}^{r}(Y_{ijl} - \overline{Y})^2$	abr-1		

Em que $\overline{Y}$, $\overline{Y}_{ij}$, $\overline{Y}_i$ e $\overline{Y}_j$ representam, respetivamente, a média geral da amostra global, a média por amostra e as médias de cada um dos níveis dos fatores A e B.
Assim, para um p-*value* inferior ao nível de significância, há motivos para rejeitar cada uma das hipóteses H_0. A rejeição da hipótese inicial significa que os fatores contribuem significativamente para explicar a variabilidade existente na variável resposta. Por sua vez, a não rejeição da hipótese inicial indica que os fatores não contribuem significativamente para a variação da variável resposta.

4.2.7 Diagnóstico de colinearidade

Quando duas ou mais variáveis independentes estão fortemente correlacionadas entre si, condição designada por colinearidade, a análise do modelo de regressão ajustado pode ser extremamente confusa e desprovida de significado, fazendo desta condição (que as variáveis independentes o sejam de facto) um dos principais pressupostos a validar durante este processo (Marôco[8]).
Assim, a colinearidade pode ser diagnosticada através da análise da matriz de correlações bivariadas entre todas as variáveis da base de dados mas não existe um valor de correlação limite a partir do qual seja possível inferir a problemas na estimação do modelo entre as

variáveis independentes.

Uma outra forma de detetar a existência de colinearidade é feita recorrendo ao Fator de Inflação da Variância (*Variance Inflation Factor - VIF*) que é definido como:

$$VIF = \frac{1}{1 - R_i^2}.$$

Note-se que

$$Var(\widehat{\beta}_i) = \sigma^2 \left(\frac{1}{1 - R_i^2} \right) \times \frac{1}{\sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2}$$

sendo R_i^2 o coeficiente de determinação do modelo entre as variáveis independentes e σ^2 pode ser estimada pelo quadrado médio dos resíduos do modelo de regressão.

Assim, se a correlação múltipla entre uma das variáveis independentes e as restantes for nula, então a variância da estimativa do parametro β será reduzida e vice-versa.

Note-se que valores de VIF superiores a cinco revelam problemas com a estimação dos parâmetros β devido à presença de colinearidade.

4.2.8 Diagnóstico de *outliers* e de observações influentes

A análise dos resíduos permite identificar *outliers* e observações influentes que terão influência na estimação do modelo. Para isso, são analisados as seguintes medidas:

- Resíduos;
- Distância de Cook;
- *Dfbetas*.

Resíduos não estandardizados

Os resíduos não estandardizados ou erros são definidos como

$$e_j = y_j - \hat{y}_j = y_j - n_j \hat{\pi}_j$$

onde y_j e $\hat{y}_j$ representam o número de sucessos observados e estimados para a célula j, respetivamente, n_j o número de observações da célula j e $\hat{\pi}_j$ a probabilidade de sucesso estimada na célula j.

Resíduos estandardizados ou resíduos de Pearson

Os resíduos estandardizados ou resíduos de Pearson são definidos como

$$e'_j = \frac{e_j}{\sqrt{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}}.$$

Note-se que:

• para amostras relativamente grandes e'_j apresenta distribuição assintótica N(0,1);

• 95% dos $|e'_j|$ devem ser inferiores a 1.96 e qualquer observação superior poderá ser classificada como um *outlier* para $\alpha = 0.05$ (uma vez que o quantil da N(0,1) de ordem 0.975 é dado por $z_{0.975} = 1.96$);

• os resíduos de Pearson não são completamente estandardizados (dependem das observações e da sua influência na estimação dos coeficientes do modelo).

Resíduos de Pearson estandardizados ou "estudentizados"

Os resíduos de Pearson estandardizados ou "estudentizados" são definidos como

$$r_j = \frac{e'_j}{\sqrt{1 - h_j}}$$

onde h_j, designada como *Leverage*, é o j-ésimo elemento da diagonal da matriz $H = (X'X)^{-1}X'$ conhecida por matriz "chapéu" e avalia a influência de cada observação no ajustamento do modelo. Se esta for aproximadamente zero é pouco importante no ajustamento do modelo enquanto que se for muito próximo de um é importante.

Note-se que os residuos de Pearson estandardizados apresentam variância constante e igual a um.

Hosmer et al.[7] demonstram que os resíduos de Pearson estandardizados e a *Leverage* para além de permitirem identificar *outliers*, também permitem avaliar a influência de uma observação no ajustamento do modelo através da variação da estatística χ^2 de Pearson:

$$\Delta\chi^2_j = \frac{(e'_j)^2}{1 - h_j} = (r_j)^2.$$

Note-se que $\Delta\chi^2_j$ apresenta distribuição assimptótica $\chi^2_{(1)}$ e que para $\alpha = 0.05$ o quantil de ordem 0.95 dessa distribuição toma o valor $\chi^2_{0.95;(1)} = 3.84$. Assim, os valores de $\Delta\chi^2_j$ superiores a 3.84 indicam observações influentes no ajustamento do modelo.

Distância de Cook

A Distância de Cook é uma medida que indica a influência de uma determinada observação na estimação dos coeficientes do modelo e é definida como

$$DC_j = r_j^2 \frac{h_j}{1 - h_j}.$$

Note-se que se esta medida tomar valores superior a um então as observações a que estão associadas deverão ser consideradas influentes na estimação dos coeficientes do modelo (Hosmer et al.[7]) e todas as que sejam superiores a $4/n$, onde n representa o número total de observações, deverão ser analisadas.

DfBetas

Os *DfBetas* permitem também estudar a influência de cada observação na estimação de cada um dos coeficientes que compõe o modelo de regressão logística e são definidos como

$$DfBetas_{ij} = \hat{\beta}_i - \hat{\beta}_{i(-j)}$$

onde $\hat{\beta}_i$ é a estimativa do coeficiente de regressão ajustado a todas as observações e $\hat{\beta}_{i(-j)}$ é a estimativa do coeficiente de regressão ajustado sem a observação j.
Note-se que valores de *DfBetas* superiores a $2\sqrt{(p+1)/n}$, onde $p+1$ representa o número de coeficientes do modelo e n a dimensão da amostra, serão observações influentes e a observação j deverá ser analisada uma vez que a sua presença no modelo afeta a estimativa de β_i.

4.2.9 Avaliação do modelo de regressão logística

Matriz de Confusão

Segue-se a descrição de algumas métricas importantes para avaliar o poder preditivo dos modelos construídos, obtidas a partir da análise da matriz de confusão (que reune o número de verdadeiros positivos - VP, falsos positivos - FP, falsos negativos - FN e verdadeiros negativos - VN):

• Precisão: a proporção de previsões corretas;

$$Precisão = (VP + VN)/n^o \text{ de dados da amostra}$$

• Sensibilidade: proporção de verdadeiros positivos;

$$Sensibilidade = VP/(VP + FN)$$

• Especificidade: proporção de verdadeiros negativos;

$$Especificidade = VN/(VN + FP)$$

Curva ROC e Ponto de Corte

Após obter as estimativas dos parâmetros do modelo de regressão logística $(\beta_0, ..., \beta_p)$ é possível estimar a probabilidade $\hat{\pi}_j$ uma vez que

$$\hat{\pi}_j = \frac{e^{\beta_0+\beta_1 X_{1j}+...+\beta_p X_{pj}}}{1 + e^{\beta_0+\beta_1 X_{1j}+...+\beta_p X_{pj}}}.$$

Assim, se $\hat{\pi}_j$ for superior a 0.5 então o indivíduo j é classificado no grupo "1-sucesso"(uma vez que é mais provável possuir o atributo associado à variável dependente do que não possuí-lo). Caso contrário, é clasificado no grupo "0-insucesso"(uma vez que é mais provável não possuir o atributo associado à variável dependente do que possuí-lo). Note-se que o valor de probabilidade 0.5 (ponto de corte ou *cut-off*) poderá ser ajustado de forma a que a classificação seja mais rigorosa.

Este ponto de corte é determinado através da Curva ROC (*Receiver Operating Characteristic Curve*), a qual é definida pela relação entre a Sensibilidade ($P(\hat{Y} = 1|Y = 1)$) e 1-Especificidade ($1 - P(\hat{Y} = 0|Y = 0)$).

Note-se que a escolha do ponto de corte é baseada numa combinação ótima entre a Sensibilidade e a 1-Especificidade e este deve situar-se o mais próximo possível do canto superior esquerdo do gráfico e onde a diferença entre a sensibilidade e a especificidade é mínima logo o somatório de ambas é máximo.

A área abaixo da curva ROC varia entre zero e um e, quanto mais próxima estiver de um, maior é a capacidade do modelo para discriminar os indivíduos que apresentam a característica de interesse (sucesso), dos indivíduos que não a apresentam (insucesso).

Salienta-se, também, a importância em dividir a amostra em duas bases de dados distintas, a de treino e a de teste, sendo a primeira utilizada para ajustar o modelo e a segunda para testá-lo. Esta divisão é feita normalmente segundo a proporção 70%/30% da base de dados inicial e de aforma aleatória.

Capítulo 5

Métodos *Shrinkage*

Os métodos de regressão de *Shrinkage*, ou de contração, são aplicados quando os modelos a construir possuem muitas variáveis (relevantes e não relevantes), podendo apresentar ou não colinearidade, sendo que muitas vezes o número de variáveis excede o número de observações, ou então existe uma forte correlação entre elas.

A regressão *Ridge* é um dos métodos *Shrinkage* que tem como objetivo específico reduzir a instabilidade dos estimadores de mínimos quadrados na presença de colinearidade, a qual é traduzida por valores elevados dos erros padrão das estimativas. Assim, as estimativas dos parâmetros β_j são escolhidas de forma a minimizar a seguinte função objetivo:

$$\Sigma_{i=1}^{n}(y_i - \Sigma_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\Sigma_{j=1}^{p}\beta_j^2$$

onde y_i representa a resposta da variável dependente para o indivíduo i, x_{ij} representa a resposta da variável independente j do indivíduo i e λ uma constante de penalização denominada constante de contração ou de regularização (The Pennsylvania State University[15]).
A escolha da constante de contração λ é habitualmente feita através da análise do traço *Ridge* ou através da validação cruzada. Quanto maior for o valor selecionado para a constante de regularização maior é a contração para zero das estimativas dos parâmetros. Se $\lambda = 0$ obtêm-se os estimadores de mínimos quadrados ordinários.

O método ***Lasso*** (*least absolute shrinkage and selection operator*) é outro dos métodos *Shrinkage* que, para além de fazer contração também faz seleção de variáveis, em que as estimativas dos parâmetros β_j são escolhidas de forma a minimizar a seguinte função objetivo:

$$\Sigma_{i=1}^{n}(y_i - \Sigma_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\Sigma_{j=1}^{p}|\beta_j|$$

onde y_i representa a resposta da variável dependente para o indivíduo i, x_{ij} representa a resposta da variável independente j do indivíduo i e λ uma constante de penalização escolhida previamente (The Pennsylvania State University[15]).

O método ***Elastic Net*** é outro dos métodos *Shrinkage* e reune os dois métodos anteriores (Zou e Hastie[16]).

Capítulo 6

Tratamento dos valores omissos/em falta

Ao longo deste trabalho, foram explorados dois métodos para o tratamento dos valores omissos/em falta, o método de imputação pelo vizinho mais próximo (*Hot Deck*) e o método de imputação pelos k-vizinhos mais próximos (KNN).

O primeiro método preenche o valor em falta de um atributo de um indivíduo/objeto a partir do valor do mesmo atributo em outro indivíduo/objeto. A imputação de dados pode ser feita a partir de um indivíduo/objeto escolhido de forma aleatória ou até de um que possua atributos muitos similares.

Para aplicar este método é necessário calcular o número de bases de dados imputadas que serão necessárias, identificado por x, sendo que este poderá variar entre 2 e 10, sendo 5 a quantidade dita ideal.

Assim, Rubin mostrou, em 1987 (Rubin[11]), que

$$x = \frac{\gamma \times Efic}{1 - Efic},$$

sendo γ a taxa de informação em falta e *Efic* a eficiência de todo este processo que se deseja estar bem próxima de 1. Note-se que quanto maior for o grau de eficiência de todo este processo maior será o número de bases de dados imputadas e consequentemente mais moroso este será a nível cumputacional. Pelas Regras de Rubin, a base de dados final (com todos os valores omissos preenchidos) é obtida a partir da média das x bases de dados imputadas.

Este método mostrou-se, de facto, muito moroso e pouco robusto uma vez que ao criar valores omissos de forma propositada na base da dados inicial e comparando-a com as bases de dados imputadas verificaram-se acentuadas discrepâncias e a variabilidade geral da população não era propriamente mantida.

O método de imputação pelos k-vizinhos mais próximos (KNN) reune os k indivíduos/objetos com atributos semelhantes aos do indivíduo/objeto com valores em falta e preenche-o pela média ponderada ou mediana desse atributo nesses k vizinhos. São tomados, por defeito dez vizinhos.
Este método foi aplicado tomando onze vizinhos e imputação pela mediana para que os valores em falta fossem preenchidos por valores que já existiam na base de dados e sobre estes não fosse efetuada qualquer tipo de média aritmética (daí o número de vizinhos ser um número ímpar).

Os valores omissos que representavam aproximadamente 1,26% dos registos da base de dados inicial foram, assim, preenchidos (vejam-se as Tabelas 6.1 a 6.3).

Note-se que dado o baixo número de valores omissos, os registos dos indivíduos a que estavam associados poderiam ter sido eliminados (e ter-se-ia resolvido esta questão de uma forma muito simples) mas optou-se pela pesquisa e utilização de métodos de imputação uma vez que é uma temática muito importante no tratamento de dados.

Tabela 6.1: Valores omissos/em falta por variável

Nome da Variável	Nº valores omissos/em falta
NascimentoOutborn	64
MaeIdade	218
IdadeGestacional	-
NascimentoPeso	-
NascimentoComprimento	504
NascimentoPerimetroCefalico	439
Transporte	87
CuidadosPrenatais	-
ConcepcaoAssistida	76
CorticoidesPrenatais	-
PatologiasNaGravidez	102
TipoDeParto	63
MotivoDoParto	157
Sexo	63
Gemelar	63
RessuscitacaoOxigenio	65
RessuscitacaoInsuflador	66
RessuscitacaoEntubacaoEt	65
RessuscitacaoCompressaoCardiaca	67
RessuscitacaoAdrenalina	68
SROxigenio	12

Tabela 6.2: Continuação dos valores omissos/em falta por variável

Nome da Variável	Nº valores omissos/faltantes
SRCpap	15
SRVppni	94
SRVentilacaoOxidoNitrico	1412
SRIppv	23
SRVaf	22
SRVafni	22
SRDiasVentilacao	-
MalformacaoCongenitaMajor	-
SepsisMeningitePrecoce	-
SepsisMeningiteTardia	-
SurfactanteInicial	66
SurfactantePosterior	116
OxigenioDia28	93
CorticoidesDPC	97
DiagSdr	21
DiagPneumotorax	31
DiagPda	40
DiagNec	42
DiagPerfuracaoGi	43
PdaProfilatico	32
PdaTerapeutico	33
CirurgiaPda	38

Tabela 6.3: Continuação dos valores omissos/em falta por variável

Nome da Variável	Nº valores omissos/faltantes
CirurgiaNec	58
CirurgiaMajorOutra	60
Hpiv	23
Evhp	29
EvhpExtensao	11
DilatacaoVentricularPh	55
LpvGrau	48
ExameOftalmologicoRopGrau	5
ExameOftalmologicoRopCirurgia	2
ExameOftalmologicoPlus	2
Apgar1	68
Apgar5	68
Apgar10	449

Capítulo 7

Análise da base de dados

A base de dados inicial deste trabalho é composta por registos de 7506 indivíduos relativamente a 128 variáveis.
As variáveis estão associadas a registos de:

• Identificação: data de nascimento, idade da mãe, nome da unidade de saúde, entre outras;

• Pré-admissão: idade gestacional, peso e comprimento ao nascer, perímetro cefálico, entre outras;

• Sala de Partos: tipo de parto, sexo, gemelar, ocorrência de ventilação, entre outras;

• Internamento: ocorrência de algum tipo de cirurgia, ocorrência de ventilação, deteção de Sépsis/Meningite precoce ou tardia, entre outras;

• Destino: data do óbito, realização de autópsia, registo de alguma complicação (respiratória, digestiva ou neurológica), entre outras.

Segue-se uma breve descrição/justificação das várias decisões que foram tomadas ao longo do trabalho:

• Não foram consideradas 72 variáveis (relacionadas com datas, registos associados a Transferência, Internamento, Óbito, entre outros) uma vez que não teriam à partida qualquer impacto nos modelos que seriam construídos, dada a parca informação por elas transmitida (elevadíssimo número de valores omissos); vejam-se as várias tabelas em apêndice com a descrição das variáveis que foram alvo de estudo;

• As variáveis resposta "Sépsis/Meningite precoce" e "Sépsis/Meningite tardia", ambas dicotómicas, foram recodificadas numa fase inicial: as respostas "Sim" e "Não" foram substituidas por "1" e "0", respetivamente;

• Criação da variável "Apgar" que traduz a existência ou ausência do indíce Apgar<6 aos 5 minutos (uma vez que, como já foi referido anteriormente, este é um fator de risco referenciado na bibliografia);

• Recodificação dos zeros da variável associada à idade da mãe para valores omissos (uma vez que estes zeros refletiam erros aquando do registo);

• Criação de duas novas variáveis dicotómicas associadas à reanimação na sala de partos e após a mesma que permitiu condensar a informação de 12 variáveis, facilitando, também, a intrepretação dos modelos; durante este processo foram criadas 7 variáveis auxiliares que foram posteriormente descartadas;

• Eliminação dos registos dos três indivíduos com sexo indeterminado (foram retirados dada a ambiguidade de serem ou não hermafroditas);

• Fusão, numa fase posterior, das duas variáveis resposta “Sépsis/Meningite precoce” e “Sépsis/Meningite tardia” numa única variável – “Sépsis/Meningite” (uma vez que particularizar o tipo de Sépsis, precoce ou tardia, deixou de ser um dos objetivos deste trabalho);

• Recodificação da variável ”SRDiasVentilacao”: os registos ”-999” (que corresponderão a indivíduos que não tiveram reanimação) tomaram o valor ”0” e a todos os outros registos foi adicionada uma unidade, de forma a ser possível diferenciar os indivíduos que foram sujeitos a reanimação dos restantes;

• Deteção de algumas incoerências nas recodificações efetuadas pelos colaboradores da empresa Mhii, associadas aos registos das variáveis “Cuidados Pré-natais”, “Malformação congénita major” e “Hpiv” (esses registos foram substituídos por valores omissos);

• Eliminação dos registos associados a um dos indivíduos, que tinha “Desconhecido” como resposta à variável “SurfactantePosterior” e era o único.

Capítulo 8

Construção de modelos

Foram construídos inúmeros modelos através do método *Stepwise*, da análise da importância das variáveis nos diversos modelos de regressão logística múltipla construídos e da ANOVA associada aos mesmos. Serão apresentados e analisados os melhores modelos para cada uma das etapas que se seguem:

• 1ª Etapa: etapa inicial de todo este processo que conduziu aos primeiros modelos, tomando como variáveis dependentes a "SepsisMeningitePrecoce" e numa fase posterior seria também tomada a "SepsisMeningiteTardia";

• 2ª Etapa: nesta etapa todas as variáveis associadas a reanimação na sala de partos e fora dela fundiram-se em duas variáveis - "ReanimacaoSalaPartos" e "ReanimacaoAposSalaPartos", respetivamente; nesta etapa a variável "SepsisMeningitePrecoce" voltou a ser tomada como variável dependente e a variável "SepsisMeningiteTardia" seria também tomada numa fase posterior;

• 3ª Etapa: nesta etapa foi criada uma nova variável - "SepsisMeningite" a partir das variáveis "SepsisMeningitePrecoce" e "SepsisMeningiteTardia" (foi assinalada a presença de sépsis/meningite em indivíduos em que tinha sido registada a presença de pelo menos uma delas) e esta passou a ser a variável dependente; a colinearidade foi minimizada;

• 4ª Etapa: nesta etapa recorreu-se à regressão por penalização - *Ridge, Lasso e Elastic Net*;

• 5ª Etapa: nesta etapa todas as ambiguidades que foram detetadas até então relativamente à codificação de algumas variáveis foram clarificadas, obtendo-se o modelo final.

8.1 1ª Etapa

Nesta 1ª Etapa foram construídos 6 modelos que seguem:

Modelo 1.1: modelo formado por todas as 54 variáveis independentes indicadas nas tabelas em apêndice; recorde-se que nesta etapa foram utilizadas as 12 variáveis relacionadas com ventilação dentro e fora da sala de partos (não sendo, portanto, ainda utilizadas as variáveis "ReanimacaoSalaPartos" e "ReanimacaoAposSalaPartos"); este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.000018 < 0.05$ - nível de significância usual);

Modelo 1.2: modelo formado por 22 variáveis independentes indicadas pelo método *Stepwise* aplicado ao modelo anterior; este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.034 <$ 0.05 - nível de significância usual);

Modelo 1.3: modelo formado pelas 6 variáveis independentes indicadas no modelo anterior com p-*value* associado ao teste de Wald menor do que o nível de significância usual (0.05); as características deste modelo encontram-se descritas nas tabelas 8.1.1 e 8.1.2; embora a área abaixo da curva ROC seja minimamente aceitável, o valor de corte é muito baixo assim como as previsões dadas por este modelo, não sendo, portanto, o modelo mais adequado;

Modelo 1.4: modelo formado pelas 18 variáveis independentes indicadas na análise da variância (ANOVA) do modelo 1.1 cujos p-*values* que estão associados aos diversos testes de hipóteses são muito reduzidos e inferiores ao nível de significância usual (0.05), contribuindo significativamente para explicar a variabilidade da variável dependente; este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.023 < 0.05$ - nível de significância usual);

Modelo 1.5: modelo formado pelas 23 variáveis independentes indicadas na análise da variância (ANOVA) do modelo 1.1 cujos p-*values* que estão associados aos diversos testes

de hipóteses são reduzidos e inferiores ao nível de significância usual (0.05), contribuindo significativamente para explicar a variabilidade da variável dependente; este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = 0.034 < 0.05 - nível de significância usual);

Modelo 1.6: modelo formado por 15 das variáveis independentes que são referidas em Hornik et al.[6], artigo que esteve na base de todo este trabalho; este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = 0.023 < 0.05 - nível de significância usual).

Tabela 8.1.1: Significância dos parâmetros do melhor modelo da 1ª Etapa

Variáveis	$\widehat{\beta}_i$	Significância	ANOVA	VIF
Constante	-4.0395	< 2e-16		
PatologiasNaGravidez: Sim	0.5642	3.56e-08	4.018e-05	1.015238
DiagSdr: Não aplicável	-12.8004	0.964249	< 2.2e-16	1.032792
DiagSdr: Sim	1.4773	< 2e-16	< 2.2e-16	1.032792
DiagNec: Sim	0.4516	0.003413	8.346e-06	1.040117
Hpiv: 1	0.5832	4.01e-06	1.104e-15	1.189241
Hpiv: 2	0.6120	3.79e-05	1.104e-15	1.189241
Hpiv: 3	0.8561	2.79e-09	1.104e-15	1.189241
Hpiv: Desconhecido	2.3836	0.001313	1.104e-15	1.189241
Hpiv: Não aplicável	-0.1712	0.580300	1.104e-15	1.189241
LpvGrau: 1	0.5501	2.66e-05	7.085e-05	1.181164
LpvGrau: 2	0.8175	0.000582	7.085e-05	1.181164
LpvGrau: 3	-0.1047	0.793219	7.085e-05	1.181164
LpvGrau: 4	0.1863	0.676051	7.085e-05	1.181164
LpvGrau: Sem informação	0.3795	0.128466	7.085e-05	1.181164
SRVafni: Sim	1.511	0.000684	0.001348	1.003248

Tabela 8.1.2: Medidas de qualidade do ajustamento do melhor modelo da 1ª Etapa

AIC	3336.4
Área abaixo da curva ROC	0.7076
Valor de corte (Cut-off)	0.12
Precisão	0.721
Sensibilidade	0.7479
Especificidade	0.4915
p-value do teste H.L.	0.5606
Nº variáveis independentes	6

8.2 2ª Etapa

Nesta 2ª Etapa foram construídos os 6 modelos que se seguem, utilizadas as 2 variáveis independentes "ReanimacaoSalaPartos" e "ReanimacaoAposSalaPartos" em detrimento das 12 variáveis "isoladas" relacionadas com ventilação dentro e fora da sala de partos; a variável "SespsisMeningitePrecoce" é a variável dependente de todos os modelos desta etapa.

Modelo 2.1: modelo formado por todas as 44 variáveis independentes indicadas nas tabelas em apêndice (exceto as que foram referidas anteriormente); este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.0087 < 0.05$ - nível de significância usual);

Modelo 2.2: modelo formado por 15 variáveis independentes indicadas pelo método *Stepwise* aplicado ao modelo anterior; este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.006 < 0.05$ - nível de significância usual);

Modelo 2.3: modelo formado pelas 3 variáveis independentes indicadas no modelo 1.1 com os menores p-*values* associados ao teste de Wald menores do que o nível de significância usual (0.05); este modelo foi rapidamente descartado uma vez que a área abaixo da cruva ROC apresentava o valor 0.6969 (e espera-se que este seja acima de 0.7) embora apresentasse um bom ajustamento aos dados em questão (p-*value* do teste de ajustamento = $0.53 > 0.05$ - nível de significância usual);

Modelo 2.4: modelo formado pelas 13 variáveis independentes indicadas na análise da variância (ANOVA) do modelo 2.1 cujos p-*values* que estão associados aos diversos testes de hipóteses são inferiores ao nível de significância usual (0.05), contribuindo significativamente para explicar a variabilidade da variável dependente "SepsisMeningitePrecoce"; as características deste modelo encontram-se descritas nas tabelas 8.2.1 e 8.2.2 uma vez que se trata do melhor modelo construído nesta etapa (embora o modelo seguinte possua características muito semelhantes); o modelo apresenta um bom ajustamento aos dados em questão e a área

da curva ROC é razoável; o valor de corte continua baixo, logo as previsões do modelo não serão as mais desejáveis; o modelo apresenta alguma colinearidade (vejam-se VIF's).

Tabela 8.2.1: Significância dos parâmetros do melhor modelo da 2ª Etapa

Variáveis	$\widehat{\beta_i}$	Significância	ANOVA	VIF
Constante	-1.579e+01	0.994750		
PatologiasNaGravidez: Sim	5.540e-01	1.73e-07	5.925e-05	1.073226
Gemelar: Sim	-1.536e-01	0.145196	2.417e-05	1.047546
NascimentoComprimento	-2.709e-02	0.246562	$<$ 2.2e-16	3.733245
NascimentoPerimetroCefalico	-1.750e-02	0.622186	5.254e-05	4.484552
MalformacaoCongenitaMajor: Não aplicável	-1.257e+01	0.964335	2.233e-06	1.012708
MalformacaoCongenitaMajor: Sim	-5.876e-01	0.045079	2.233e-06	1.012708
ReanimacaoAposSalaPartos	7.968e-01	0.023584	2.156e-12	1.185914
SurfactantePosterior: Não	1.351e+01	0.995509	$<$ 2.2e-16	1.479981
SurfactantePosterior: Sim	1.446e+01	0.995193	$<$ 2.2e-16	1.479981
DiagSdr: Sim	6.526e-01	0.000347	0.0001133	1.451908
Sexo: Masculino	1.644e-01	0.081219	0.0519825	1.032590
Hpiv: 1	5.268e-01	4.03e-05	4.354e-07	1.313376
Hpiv: 2	4.780e-01	0.001648	4.354e-07	1.313376
Hpiv: 3	5.410e-01	0.000353	4.354e-07	1.313376
Hpiv: Desconhecido	2.370e+00	0.001639	4.354e-07	1.313376
Hpiv: Não aplicável	-2.663e-01	0.401287	4.354e-07	1.313376
IdadeGestacional:	-1.609e-03	0.736881	0.7385241	3.578288
MaeIdade:	-1.310e-02	0.092293	0.0934992	1.034153
Apgar10:	1.096e-03	0.455237	0.4585522	1.005728

Tabela 8.2.2: Medidas de qualidade do ajustamento do melhor modelo da 2ª Etapa

AIC	3258
Área abaixo da curva ROC	0.7351
Valor de corte (Cut-off)	0.12
Precisão	0.6793
Sensibilidade	0.6779
Especificidade	0.6907
p-value do teste H.L.	0.2355
Nº variáveis independentes	13

Modelo 2.5: modelo formado pelas 5 variáveis independentes indicadas no modelo 2.1 com os menores p-*values* associados ao teste de Wald menores do que o nível de significância usual (0.05); as medidas de qualidade do ajustamente deste modelo são relativamente inferiores às do modelo anterior embora se encontre melhor ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.605 > 0.05$ - nível de significância usual).

Modelo 2.6: modelo formado por 11 das variáveis independentes que são referidas em Hornik et al.[6], artigo que esteve na base de todo este trabalho; este modelo foi rapidamente descartado uma vez que embora se encontrasse ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.325 > 0.05$), a área abaixo da curva ROC era inferior a 0.7.

8.3 3ª Etapa

Nesta 3ª Etapa foram construídos os 5 modelos que se seguem e utilizadas as 2 variáveis independentes "ReanimacaoSalaPartos" e "ReanimacaoAposSalaPartos" em detrimento das 12 variáveis "isoladas" relacionadas com ventilação dentro e fora da sala de partos; a variável "SespsisMeningite" é a variável dependente de todos os modelos desta etapa; nesta etapa a colinearidade foi minimizada através da exclusão de variáveis que seriam linearmente dependentes (através do comando *alias*) e dos *VIF*'s, melhorando, assim, as probabilidade de deteção de presença e ausência das infeções em questão:

Modelo 3.1: modelo formado apenas por 27 variáveis independentes indicadas nas tabelas em apêndice; numa fase inicial foi detetada a existência das variáveis que se seguem que seriam linearmente dependentes através do comando *alias* e que estariam, assim, a perturbar as previsões dadas pelo modelo (tendo sido excluídas):

- ExameOftalmologicoRopCirurgia
- MalformacaoCongenitaMajor
- ExameOftalmologicoRopGrau
- LpvGrau
- CorticoidesDPC
- DiagPneumotorax
- DiagPda
- DiagNec
- DiagPerfuracaoGi
- PdaProfilatico
- PdaTerapeutico
- CirurgiaPda
- CirurgiaNec
- CirurgiaMajorOutra

Foram detetados sucessivamente os seguintes VIF's associados às variáveis apresentadas na tabela 8.3.1.

Tabela 8.3.1: VIF's da variáveis independentes que não foram incluídas no modelo 3.1

Variáveis	VIF
Transporte	45291500
Evhp	17140.316335
DilatacaoVentricularPh	90.678973

As medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.2.2. Há melhorias significativas no valor de corte (comparativamente com as duas etapas anteriores) logo as previsões dadas pelo modelo tornaram-se mais aceitáveis. O AIC é outra das medidas que se destaca do dos modelos construídos nas etapas anteriores, tendo aumentado consideravelmente.

Modelo 3.2: modelo formado por 17 variáveis independentes indicadas pelo método *Stepwise* aplicado ao modelo anterior; as medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.2.2;

Modelo 3.3: modelo formado pelas 14 variáveis independentes indicadas na análise da variância (ANOVA) do modelo anterior cujos p-*values* que estão associados aos diversos testes de hipóteses são muito reduzidos e inferiores ao nível de significância usual (0.05); as características deste modelo encontram-se descritas nas tabelas 8.3.2 e 8.3.3, uma vez que foi o melhor modelo construído nesta etapa; os VIF's de todas as variáveis são inferiores a 5, o que revela a quase inexistência de colinearidade; o AIC é bastante elevado (comparativamento ao dos modelos apresentados nas etapas anteriores) mas praticamente todas as restantes medidas foram melhoradas; o valor de corte aumentou consideravelmente, encontrando-se bem próximo de 0.5 que é o desejável; este modelo encontra-se bem ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.3697 > 0.05$) ;

Modelo 3.4: modelo formado pelas 16 variáveis independentes indicadas na análise da variância (ANOVA) do modelo 1.1 cujos p-*values* que estão associados aos diversos testes de hipóteses são muito reduzidos e inferiores ao nível de significância usual (0.05), contribuindo significativamente para explicar a variabilidade da variável dependente; as características deste modelo são muito semelhantes às dos anteriores embora não esteja tão bem ajustado

(p-*value* do teste de ajustamento = 0.1305 > 0.05 - nível de significância usual);

Tabela 8.3.2: Significância dos parâmetros do melhor modelo da 3ª Etapa

Variáveis	$\widehat{\beta}_i$	Significância	ANOVA	VIF
Constante	-1.343e+01	0.992636		
CorticoidesPrenatais: Desconhecido	-6.371e-01	0.280505	1.269e-08	1.141332
CorticoidesPrenatais: Não	-3.233e-01	0.008301	1.269e-08	1.141332
CorticoidesPrenatais: Parcial	-2.427e-01	0.002707	1.269e-08	1.141332
PatologiasNaGravidez: Sim	2.198e-01	0.003559	7.688e-06	1.181904
Gemelar: Sim	-1.795e-01	0.011950	5.409e-10	1.053992
NascimentoPerimetroCefalico	5.132e-02	0.047152	< 2.2e-16	3.835136
TipoDeParto: Vaginal	-2.474e-01	0.002576	3.078e-06	1.246661
Sexo: Masculino	1.046e-01	0.115712	8.665e-05	1.041605
ReanimacaoAposSalaPartos: Sim	8.720e-01	3.21e-08	< 2.2e-16	1.204658
SurfactantePosterior: Não	1.471e+01	0.991936	< 2.2e-16	1.561305
SurfactantePosterior: Sim	1.502e+01	0.991766	< 2.2e-16	1.561305
DiagSdr: Não aplicável	3.959e+01	0.810147	2.247e-07	1.509425
DiagSdr: Sim	1.911e-01	0.037441	2.247e-07	1.509425
Hpiv: 1	2.242e-01	0.023641	< 2.2e-16	1.374923
Hpiv: 2	1.734e-01	0.174319	< 2.2e-16	1.374923
Hpiv: 3	-1.340e-01	0.309860	< 2.2e-16	1.374923
Hpiv: Desconhecido	1.451e+00	0.090570	< 2.2e-16	1.374923
Hpiv: Não aplicável	-1.876e+00	1.36e-10	< 2.2e-16	1.374923
ExameOftalmologicoPlus: Não aplicável	-5.435e-01	0.000327	3.973e-15	1.122628
ExameOftalmologicoPlus: Sim	-3.272e-01	0.359956	3.973e-15	1.122628
IdadeGestacional	-1.215e-02	0.000476	4.596e-11	3.268311
NascimentoPeso	-1.121e-03	9.87e-10	2.801e-11	3.399007
SRDiasVentilacao	5.178e-02	< 2e-16	< 2.2e-16	1.371562

Tabela 8.3.3: Medidas de qualidade do ajustamento do melhor modelo da 3ª Etapa

AIC	5651.8
Área abaixo da curva ROC	0.7994
Valor de corte (Cut-off)	0.43
Precisão	0.7437
Sensibilidade	0.8159
Especificidade	0.6222
p-value do teste H.L.	0.3697
Nº variáveis independentes	14

Modelo 3.5: modelo formado por 7 das variáveis independentes indicadas no modelo 3.1 com os menores p-*values* associados ao teste de Wald menores do que o nível de significância usual (0.05); este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = 0.02875 < 0.05 - nível de significância usual);

8.4 4ª Etapa

Nesta 4ª Etapa foram construídos os 3 modelos que seguem, utilizadas as 2 variáveis independentes "ReanimacaoSalaPartos" e "ReanimacaoAposSalaPartos" em detrimento das 12 variáveis "isoladas" relacionadas com ventilação dentro e fora da sala de partos. A variável "SespsisMeningite" é a variável dependente de todos os modelos desta etapa e foi utilizada a regressão com métodos de contração de *Shrinkage*, ou de penalidade, tomando o $\lambda_{mínimo}$. Esta temática não foi muito desenvolvida uma vez que não conduziu a melhores resultados. Assim, a regressão por penalidade pode ser aplicada, por exemplo, segundo os seguintes métodos:

- *Lasso*
- *Ridge*
- *Elastic Net*

Modelo 4.1: modelo formado pelas 14 variáveis independentes que constituem o modelo 3.2 (melhor modelo da 3ª etapa) com recurso à regressão por penalidade - *Lasso*; as características deste modelo são as que constam nas tabelas 8.4.1 e 8.4.2; comparando as medidas de qualidade do ajustamento deste modelo com as do melhor modelo da etapa anterior verifica-se que não ocorreram melhorias, razão pelo qual a regressão por penalidade não foi muito aprofundada;

Modelo 4.2: modelo formado pelas 14 variáveis independentes que constituem o modelo 3.2 (melhor modelo da 3ª etapa) com recurso à regressão por penalidade - *Ridge*; este modelo foi rapidamente descartado uma vez que não se encontrava ajustado aos dados em questão (p-*value* do teste de ajustamento = $0.003 < 0.05$ - nível de significância usual);

Modelo 4.3: modelo formado pelas 14 variáveis independentes que constituem o modelo 3.2 (melhor modelo da 3ª etapa) com recurso à regressão por penalidade - *Elastic Net*; as medidas de qualidade do ajustamento deste modelo não diferem muito das do primeiro modelo embora não esteja tão bem ajustado aos dados (p-*value* do teste de ajustamento = $0.2146 > 0.05$ - nível de significância usual).

Tabela 8.4.1: Significância dos parâmetros do melhor modelo da 4ª Etapa

Variáveis	$\widehat{\beta}_i$
Constante	1.44719326
. CorticoidesPrenatais: Desconhecido	-0.56975601
CorticoidesPrenatais: Não	-0.31383525
CorticoidesPrenatais: Parcial	-0.23636293
PatologiasNaGravidez: Sim	0.21969772
Gemelar: Sim	-0.17396196
NascimentoPerimetroCefalico	0.04428937
TipoDeParto: Vaginal	-0.24734979
Sexo: Masculino	0.10497031
ReanimacaoAposSalaPartos: Sim	0.86363091
SurfactantePosterior: Sim	0.32129147
DiagSdr: Sim	0.19060359
Hpiv: 1	0.22485541
Hpiv: 2	0.18217197
Hpiv: 3	-0.11680847
Hpiv: Desconhecido	1.41900389
Hpiv: Não aplicável	-1.86754394
ExameOftalmologicoPlus: Não aplicável	-0.55068756
ExameOftalmologicoPlus: Sim	-0.25843578
IdadeGestacional	-0.01209388
NascimentoPeso	-0.00110062
SRDiasVentilacao	0.04689382

Tabela 8.4.2: Medidas de qualidade do ajustamento do melhor modelo da 4ª Etapa

Área abaixo da curva ROC	0.7988
Valor de corte (Cut-off)	0.423
Precisão	0.7437
Sensibilidade	0.8045
Especificidade	0.6412
p-value do teste H.L.	0.3314
Nº variáveis independentes	14

8.5 5ª Etapa

Nesta 5ª Etapa foram construídos os 6 modelos que seguem, utilizando as 2 variáveis independentes "ReanimacaoSalaPartos" e "ReanimacaoAposSalaPartos" em detrimento das 12 variáveis "isoladas" relacionadas com ventilação dentro e fora da sala de partos.
Foram retificadas todas as ambiguidades ao nível da codificação de algumas variáveis:
• os registos com o valor "-999" para a variável "SRDiasVentilacao" passaram a ser representados pelo valor 0 e a todos os outros foi adicionada uma unidade;
• a variável "Apgar10" foi ignorada uma vez que possuía 972 registos com "99" e só deveria comtemplar registos entre 1 e 10;
• foram excluídos os registos de 57 indivíduos cujas respostas às variáveis "Apgar1", "SRDiasVentilaca" e "Apgar5" eram -999" (sem informação);
• foi excluído o único registo de um indivíduo cuja resposta à variável "SurfactantePosterior" era "Desconhecido", porque sendo apenas um caso isolado não iria treinar o modelo convenientemente.
O melhor modelo construído nesta última abordagem está na base da construção da aplicação no Shiny e, comparando as suas medidas de qualidade do ajustamento com as dos modelos iníciais, há uma evolução positiva.
A variável "SespsisMeningite" é a variável dependente de todos os modelos desta etapa :

Modelo 5.1: modelo formado apenas por 26 variáveis independentes das que são indicadas nas tabelas em apêndice (exceto as que foram referidas anteriormente);numa fase inicial foi detetada a existência das variáveis que se seguem que seriam linearmente dependentes através do comando *alias* assim como variáveis com elevados VIF's (veja-se tabela 8.5.1) e que estariam, assim, a perturbar as previsões dadas pelo modelo (e foram excluídas):
• CorticoidesDPC
• DiagSdr
• DiagPneumotorax
• DiagPda
• DiagNec
• DiagPerfuracaoGi

- PdaProfilatico
- PdaTerapeutico
- CirurgiaPda
- CirurgiaNec
- CirurgiaMajorOutra
- LpvGrau
- ExameOftalmologicoPlus

Tabela 8.5.1: VIF's da variáveis independentes que não foram incluídas no modelo 3.1

Variáveis	VIF
Transporte	106356400
Evhp	5024.5
ExameOftalmologicoRopGrau	258.7
DilatacaoVentricularPh	74.4

As medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.5.2 mas este modelo não se encontra tão bem ajustado aos dados em questão quanto o melhor modelo desta etapa (p-*value* do teste de ajustamento = 0.096 > 0.05 - nível de significância usual);

Modelo 5.2: modelo formado por 17 variáveis independentes indicadas pelo método *Stepwise* aplicado ao modelo anterior; as medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.5.2 mas este modelo não se encontra tão bem ajustado aos dados em questão quanto o melhor modelo desta etapa (p-*value* do teste de ajustamento = 0.2158 > 0.05 - nível de significância usual);

Modelo 5.3: modelo formado pelas 15 variáveis independentes indicadas na análise da variância (ANOVA) do modelo anterior cujos p-*values* que estão associados aos diversos testes de hipóteses são muito reduzidos e inferiores ao nível de significância usual (0.05), contribuindo significativamente para explicar a variabilidade da variável dependente; as medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.5.2 mas este modelo não se encontra tão bem ajustado aos dados em questão quanto o melhor modelo desta etapa (p-*value* do teste de ajustamento = 0.1436 > 0.05 - nível de significância usual);

Modelo 5.4: modelo formado pelas 16 variáveis independentes indicadas na análise da variância (ANOVA) do modelo 1.1 cujos p-*values* que estão associados aos diversos testes de hipóteses são muito reduzidos e inferiores ao nível de significância usual (0.05), contribuindo significativamente para explicar a variabilidade da variável dependente; as medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.5.2 mas este modelo não se encontra tão bem ajustado aos dados em questão quanto o melhor modelo desta etapa (p-*value* do teste de ajustamento = $0.1348 > 0.05$ - nível de significância usual);

Modelo 5.5: modelo formado pelas 9 variáveis independentes indicadas no modelo 5.1 com os menores p-*values* associados ao teste de Wald menores do que o nível de significância usual (0.05); as medidas de qualidade do ajustamento deste modelo são muito semelhantes às apresentadas na tabela 8.5.2 mas este modelo não se encontra tão bem ajustado aos dados em questão quanto o melhor modelo desta etapa (p-*value* do teste de ajustamento = $0.1128 > 0.05$ - nível de significância usual);

Modelo 5.6: modelo formado por 8 das variáveis independentes indicadas no modelo 5.2 com os menores p-values associados ao teste de Wald menores do que o nível de significância usual (0.05); as características deste modelo encontram-se descritas nas tabelas 8.5.1 e 8.5.2; a área abaixo da curva ROC é aceitável e o valor de corte encontra-se muito próximo de 0.5, logo as previsões dadas por este modelo são bastante aceitáveis;
Note-se que, embora algumas variáveis possuam o p-*value* associado ao teste de Wald maior do que o nível de significância usual (0.05), o p-*value* das mesmas associado ao teste de hipóteses na análise da variância (ANOVA) é sempre inferior ao nível de significância usual (0.05), logo não foram descartadas na construção do modelo até porque são variáveis com poder explicativo na váriavel dependente em questão.

Tabela 8.5.2: Significância dos parâmetros do melhor modelo da 5ª Etapa

Variáveis	$\widehat{\beta}_i$	Significância	ANOVA	VIF
Constante	1.8464735	0.004908		
CorticoidesPrenatais: Desconhecido	-0.8308289	0.137754	2.666e-12	1.095553
CorticoidesPrenatais: Não	-0.4113552	0.000782	2.666e-12	1.095553
CorticoidesPrenatais: Parcial	-0.2892865	0.000315	2.666e-12	1.095553
MotivoDoParto: IVG	0.4010217	3.22e-06	0.0003841	1.314303
MotivoDoParto: Patologia Fetal	0.3761059	2.03e-05	0.0003841	1.314303
MotivoDoParto: Patologia materna	0.3494295	0.572000	0.0003841	1.314303
ReanimacaoAposSalaPartos: Sim	1.1349879	1.20e-13	< 2.2e-16	1.093008
SurfactantePosterior: Sim	0.3137491	4.95e-05	< 2.2e-16	1.363520
Hpiv: 1	0.2658478	0.007586	< 2.2e-16	1.366423
Hpiv: 2	-0.0159248	0.901781	< 2.2e-16	1.366423
Hpiv: 3	-0.1902289	0.155177	< 2.2e-16	1.366423
Hpiv: Desconhecido	1.2135156	0.104283	< 2.2e-16	1.366423
Hpiv: Não aplicável	-2.0635755	1.94e-12	< 2.2e-16	1.366423
IdadeGestacional	-0.0141152	3.34e-05	< 2.2e-16	3.051207
NascimentoPeso	-0.0008060	5.00e-08	4.694e-13	2.209854
SRDiasVentilacao	0.0628684	< 2e-16	< 2.2e-16	1.300047

Tabela 8.5.3: Medidas de qualidade do ajustamento do melhor modelo da 5ª Etapa

AIC	5556.6
Área abaixo da curva ROC	0.7844
Valor de corte (Cut-off)	0.42
Precisão	0.7332
Sensibilidade	0.8060
Especificidade	0.6106
p-value do teste H.L.	0.2908
Nº variáveis independentes	8

Procedeu-se à análise dos resíduos e *outliers* do modelo 5.6 - melhor modelo de todas as etapas apresentadas, fazendo a representação gráfica dos resíduos (Figura 8.1), dos resíduos de Pearson (Figura 8.2) e das distâncias de Cook (Figura 8.3).

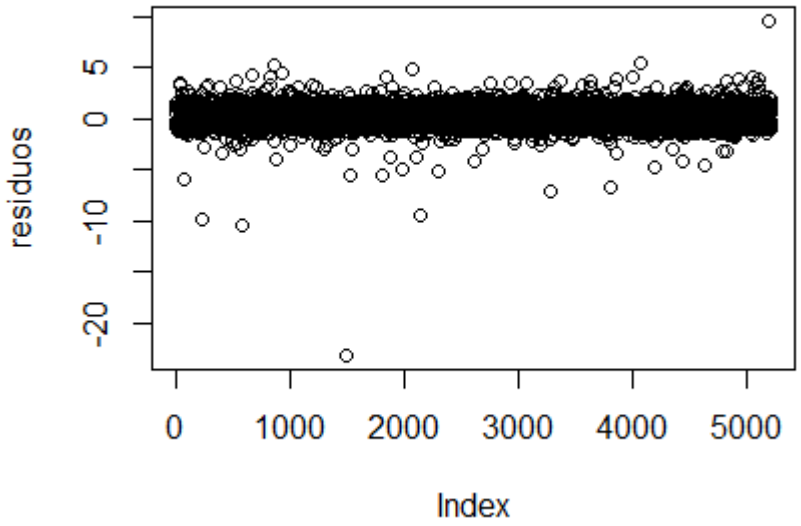


Fig. 8.1: Representação gráfica dos resíduos de Pearson do modelo 5.6

Os resíduos de Pearson apresentam média=-0.0184885 e variância=1.173819, bem próximos de 0 e 1, respetivamente, como é desejável. Note-se que 95% dos resíduos são inferiores em valor absoluto a 1.96.

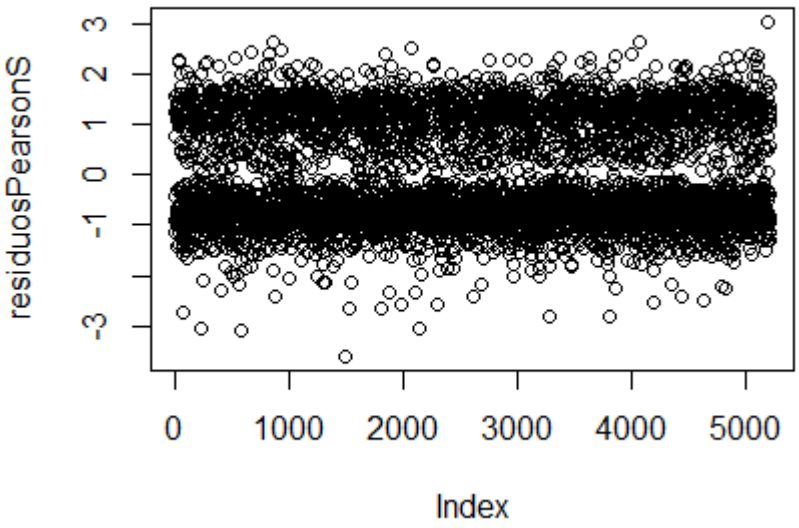


Fig. 8.2: Representação gráfica dos resíduos de Pearson "Estudentizados" do modelo 5.6

Os resíduos de Pearson "Estudentizados" apresentam média=–0.07171118 e variância=1.058203, bem próximos de 0 e 1, respetivamente, como é desejável.

Procedeu-se seguidamente à análise dos *outliers* do modelo 5.6 - melhor modelo de todas as etapas apresentadas.

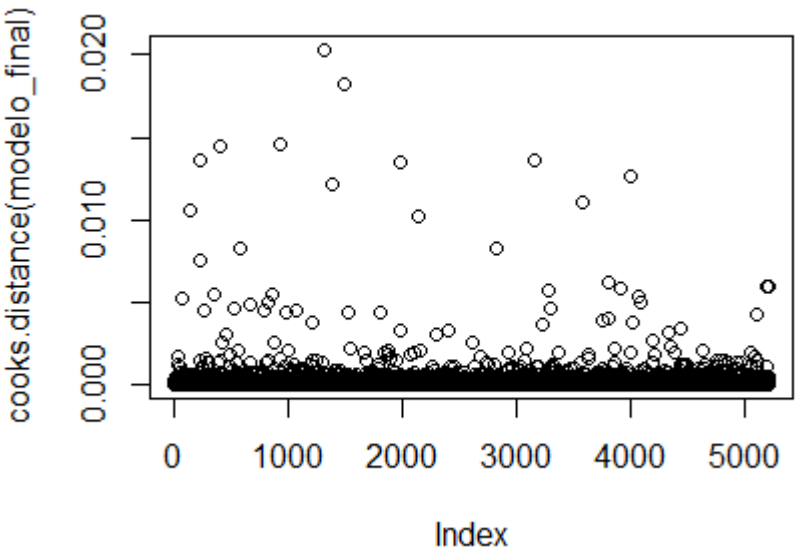


Fig. 8.3: Representação gráfica das distâncias de Cook do modelo 5.6

Analisando a representação gráfica anterior verifica-se que será importante analisar o impacto que a exclusão das observações identificadas pelos pontos mais dispersos na representação gráfica anterior poderá causar no ajustamento do modelo. Esta análise será analisada em três etapas, mediante a seguinte divisão:

•distância de Cook<0.01 - cor amarela;

•distância de Cook<0.005 - cor verde;

•distância de Cook<4/5212 - cor vermelha.

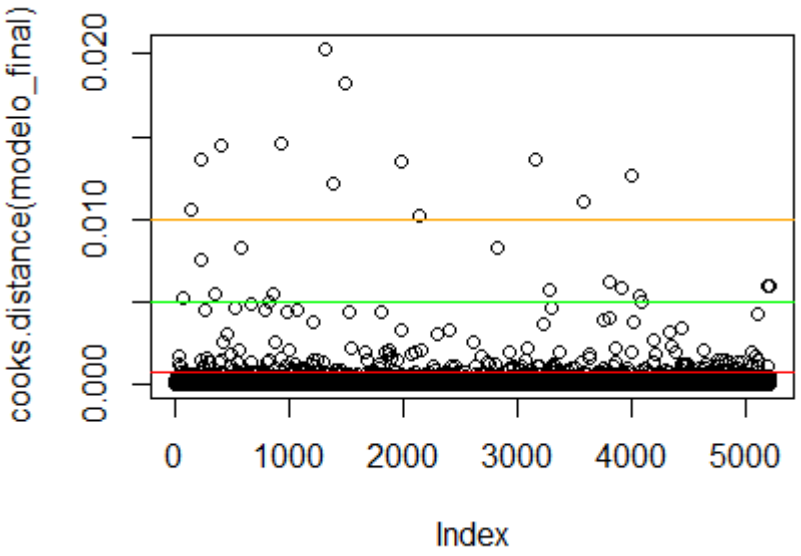


Fig. 8.4: Representação gráfica das distâncias de Cook divididas do modelo 5.6

Tabela 8.5.4: Significância dos parâmetros associados às variáveis que compõem o modelo 5.6, tendo em conta as d_{Cook}

Variáveis	$\widehat{\beta}_i$ $d_{Cook} < 0.01$	$\widehat{\beta}_i$ $d_{Cook} < 0.005$	$\widehat{\beta}_i$ $d_{Cook} < 4/5212$
Constante	1.725	1.737	1.732
CorticoidesPrenataisDesconhecido	-1.328e+01	-1.329e+01	-1.327e+01
CorticoidesPrenataisNão	-4.279e-01	-4.309e-01	-4.433e-01
CorticoidesPrenataisParcial	-2.874e-01	-2.966e-01	-2.760e-01
MotivoDoPartoIVG	3.999e-01	4.040e-01	4.179e-01
MotivoDoPartoPatologia Fetal	3.755e-01	3.832e-01	4.205e-01
MotivoDoPartoPatologia materna	4.052e-01	3.998e-01	4.069e-01
ReanimacaoAposSalaPartosSim	1.125	1.126	1.182
SurfactantePosteriorSim	3.094e-01	3.024e-01	3.116e-01
Hpiv1	2.580e-01	2.472e-01	2.529e-01
Hpiv2	-2.160e-02	-3.251e-02	-4.804e-02
Hpiv3	-2.078e-01	-2.134e-01	-2.284e-01
HpivDesconhecido	2.277	2.270	2.287
HpivNão aplicável	-1.992	-1.991	-2.027
IdadeGestacional	-1.363e-02	-1.367e-02	-1.399e-02
NascimentoPeso	-7.960e-04	-7.989e-04	-7.978e-04
SRDiasVentilacao	6.875e-02	7.059e-02	6.999e-02

Tabela 8.5.5: Medidas de qualidade do ajustamento dos modelos caracterizados na tabela 8.5.4

	d_{Cook}<0.01	d_{Cook}<0.005	d_{Cook}<4/5212
AIC	5499.5	5466.5	5252.5
Área abaixo da curva ROC	0.7817	0.7816	0.7813
Valor de corte (Cut-off)	0.424	0.42	0.421
Precisão	0.7346	0.7332	0.7323
Sensibilidade	0.8146	0.8096	0.8074
Especificidade	0.5998	0.6046	0.6058
p-value do teste H.L.	0.372	0.3268	0.3107
Nº variáveis independentes	8	8	8

Analisando a tabela 8.5.4 verifica-se que as estimativas dos parâmetros para cada um dos modelos não sofreram grandes oscilações (o que revela que as observações que foram sucessivamente retiradas não influenciam as estimativas dos parâmetros dos modelos) e as medidas de qualidade do ajustamento do modelo cujas observações possuem distância de Cook inferiores a 0.01 são relativamente melhores do que os restantes.

Analisando as tabelas 8.5.2 e 8.5.4 verifica-se que as estimativas dos parâmetros sofreram alterações significativas, logo as primeiras 12 observações que foram retiradas inicialmente (cujas distâncias Cook eram superiores ou iguais a 0.01) influenciam as estimativas dos parâmetros. Assim, o modelo obtido a partir do "refinamento" do modelo 5.6, considerando apenas as observações cujas distâncias Cook são inferiores a 0.01, é o que melhor se ajusta aos dados em estudo.

Todavia, verificou-se que ao retirar as 12 observações referidas anteriormente, eliminavam-se os registos dos únicos 4 indivíduos da sub-base de dados Treino (sobre a qual é construída o modelo) cujas respostas às variáveis "CorticoidesPrenatais" e "SepsisMeningite" eram simultanemanete Desconhecido e Sim, respetivamente, o que não é desejável uma vez que o modelo não seria treinado para este tipo de *input*.

Assim, retirando as restantes oito observações obteve-se um modelo cujas características são apresentadas nas tabelas 8.5.6 e 8.5.7.

Analisando as tabelas 8.5.2, 8.5.3, 8.5.6 e 8.5.7 verifica-se que as estimativas dos parâmetros para cada um dos modelos não sofreram grandes oscilações, o que revela que as oito observações que foram retiradas não têm grande influência na construção do modelo. Assim,

o modelo 5.6, cujas características são apresentadas nas tabelas 8.5.2 e 8.5.3 (área baixo da curva ROC é elevada, valor de corte muito próximo de 0.5 o que revela previsões mais satisfatórias e o modelo com bom ajustamento aos dados em questão) é o modelo que será utilizado para construir a aplicação *web* descrita no capítulo seguinte.

Tabela 8.5.6: Significância dos parâmetros associados ao modelo 5.6 sem 8 observações

Variáveis	$\widehat{\beta_i}$	Significância	ANOVA
Constante	1.6683764	0.011460	
CorticoidesPrenatais: Desconhecido	-0.8292963	0.138552	1.987e-12
CorticoidesPrenatais: Não	-0.4273803	0.000526	1.987e-12
CorticoidesPrenatais: Parcial	-0.2874335	0.000363	1.987e-12
MotivoDoParto: IVG	0.3959335	4.66e-06	0.000395
MotivoDoParto: Patologia Fetal	0.3786909	1.88e-05	0.000395
MotivoDoParto: Patologia materna	0.4084557	0.580900	0.000395
ReanimacaoAposSalaPartos: Sim	1.1392105	1.13e-13	< 2.2e-16
SurfactantePosterior: Sim	0.3131988	5.53e-05	< 2.2e-16
Hpiv: 1	0.2595508	0.009390	< 2.2e-16
Hpiv: 2	-0.0244771	0.850570	< 2.2e-16
Hpiv: 3	-0.2127904	0.114863	< 2.2e-16
Hpiv: Desconhecido	2.2833481	0.044382	< 2.2e-16
Hpiv: Não aplicável	-2.0401820	3.30e-12	< 2.2e-16
IdadeGestacional	-0.0133559	9.35e-05	< 2.2e-16
NascimentoPeso	-0.0008073	5.26e-08	3.386e-13
SRDiasVentilacao	0.0686630	< 2e-16	< 2.2e-16

Tabela 8.5.7: Medidas de qualidade do ajustamento do modelo 5.6 sem 8 observações

AIC	5518.6
Área abaixo da curva ROC	0.7848
Valor de corte (Cut-off)	0.414
Precisão	0.7323
Sensibilidade	0.8031
Especificidade	0.6130
p-value do teste H.L.	0.2893
Nº variáveis independentes	8

Relativamente aos *Dfbetas* do modelo com melhor poder preditivo (cujas caraterísticas são apresentadas nas tabela 8.5.2 e 8.5.3), verificou-se que estes são inferiores a 2 como seria, a grosso modo, desejável.

Como foi referido anteriormente, é desejável que os *Dfbetas* sejam, em rigor, inferiores a

$2\sqrt{(p+1)/n}$, onde $p+1$ representa o número de coeficientes do modelo e n a dimensão da amostra, de forma a garantir a inexistência de observações influentes.

Verificou-se a existência de 28 indivíduos cujos *Dfbetas* são superiores a $2\sqrt{(17)/5204} \simeq 0.114$, atendendo a que a amostra Treino sobre a qual são calculados os *Dfbetas* reune os registos de 5204 indivíduos e o modelo em questão possui 17 coeficientes.

Optou-se por não se verificar se os registos destes 28 indivíduos teriam ou não influência nas estimativas dos parâmetros do modelo uma vez que a análise anterior foi satisfatória.

Capítulo 9

Cálculo das previsões

Após a construção do modelo com melhor poder preditivo, as previsões do risco de Sépsis/Meningite em prematuros de muito baixo peso são calculadas mediante a relação entre as estimativas dos parâmetros obtidas com a informação desse indivíduo, apresentada anteriormente:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}.$$

Tomando um indivíduo com as seguintes características relativamente às variáveis que compõem o modelo com melhor poder preditivo construído no capítulo anterior e cujas características estão indicadas nas tabelas 8.5.2 e 8.5.3:

- "$CorticoidesPrenatais$" - Completo;
- "$MotivoDoParto$" - Espontâneo;
- "$ReanimacaoAposSalaParos$" - Sim;
- "$SurfactantePosterior$" - Sim;
- "$Hpiv$" - 0;
- "$IdadeGestacional$" - 215;
- "$NascimentoPeso$" - 1500;
- "$SRDiasVentilacao$" - 2.

Recorde-se que, como foi referido anteriormente, a variável "SRDiasVentilacao" foi recodificada: os registos "-999" (que correspondiam a indivíduos que não tiveram reanimação)

tomaram o valor ”0” e a todos os outros registos foi adicionada uma unidade, de forma a ser possível diferenciar os indivíduos que foram sujeitos a reanimação dos restantes.

Assim, é necessário ter isto em conta apenas aquando do cálculo manual do valor probabilístico que se segue (o número de dias em ventilação tomará o valor 3):

$$\hat{\pi} = \frac{e^{\left[\begin{array}{l} 1.8464735 + 0 \times CorticoidesPrenatais(Completo) + 0 \times MotivoDoParto(Espontâneo) + \\ +1.1349879 \times ReanimacaoAposSalaPartos(Sim) + 0.3137491 \times SurfactantePosterior(Sim) + \\ +0 \times Hpiv(0) - 0.0141152 \times IdadeGestacional(215) - 0.000806 \times NascimentoPeso(1500) + \\ +0.0628684 \times SRDiasVentilacao(3) \end{array}\right]}}{1 + e^{\left[\begin{array}{l} 1.8464735 + 0 \times CorticoidesPrenatais(Completo) + 0 \times MotivoDoParto(Espontâneo) + \\ +1.1349879 \times ReanimacaoAposSalaPartos(Sim) + 0.3137491 \times SurfactantePosterior(Sim) + \\ +0 \times Hpiv(0) - 0.0141152 \times IdadeGestacional(215) - 0.000806 \times NascimentoPeso(1500) + \\ +0.0628684 \times SRDiasVentilacao(3) \end{array}\right]}}. \tag{9.1}$$

Desta forma,

$$\hat{\pi} \simeq 0.319.$$

Este resultado poderá ser confirmado utilizando a aplicação cuja descrição é apresentada no capítulo seguinte e que foi desenvolvida expressamente para a Mhii. Note-se que a aplicação *web* já foi construída tendo em conta todos os ajustes que foram efetuados em termos de recodificação, pelo que os valores introduzidos pelo utilizador não deverão ser alvo de acréscimos prévios (como no cálculo manual anterior).

Capítulo 10

Aplicação *web*

O modelo apresentado na 5ª Etapa, identificado como o Modelo Final, esteve na base da construção de uma aplicação através do *Shiny*, sistema para desenvolvimento de aplicações web que utiliza o R.

Esta aplicação permite avaliar, em tempo real, o ligeiro, moderado ou elevado risco de Sépsis/Meningite em Prematuros de muito baixo peso, auxiliando assim muitíssimo a comunidade médica a diagnosticar e atuar precocemente.

A utilização desta aplicação permitirá, assim, travar o número de óbitos associados às infeções referidas anteriormente.

A aplicação é composta por diversos campos que requerem a introdução, por parte do utilizador, dos dados associados às variáveis com poder preditivo do modelo apresentado na 5ª Etapa. Assim, no campo assinalado por:

• Surfactante após sala de partos o utilizador deverá selecionar uma das seguintes opções: Sim / Não;

• Reanimação após sala de partos o utilizador deverá selecionar uma das seguintes opções: Sim / Não;

• Motivo do parto o utilizador deverá selecionar uma das seguintes opções: Espontâneo / IVG / Patologia Fetal / Patologia Materna;

• Corticoides Pré-natais o utilizador deverá selecionar uma das seguintes opções: Completo / Parcial / Desconhecido / Não;

• Pior grau Hpiv o utilizador deverá selecionar uma das seguintes opções: 0 / 1 /2 /3 / Não

aplicável / Desconhecido;

• Ventilação o utilizador deverá selecionar uma das seguintes opções: Sim / Não;

• Peso ao nascer o utilizador deverá indicar o peso do recém-nascido em gramas;

• Nº dias em Ventilação o utilizador deverá indicar o número de dias em que o recém-nascido se encontra em ventilação; note-se que este campo só surgirá caso o utilizador responda afirmativamente ao campo Ventilação

• Idade gestacional o utilizador deverá indicar a idade gestacional do recém-nascido em dias;

Após os campos estarem todos preenchidos, o utilizador deverá carregar em "Cálculo" e obterá, de imediato, a estimativa da probabilidade de risco de Sépsis/Meningite para um recém-nascido com essas características. Surgirá, também, a indicação se o risco é:

• ligeiro - estimativa da probabilidade <50%;

• moderado - estimativa da probabilidade maior ou igual a 50% e menor do que 75%;

• elevado - estimativa da probabilidade maior ou igual a 75%.

Após a devida experimentação desta aplicação, verifica-se que a variável "$SRDiasVentilacao$" e os níveis das respetivas variáveis "$Sim - SurfactantePosterior$", "$Sim - ReanimacaoAposSalaPartos$", "$Sim - Ventilacao$" e parto não espontâneo contribuem de forma positiva no aumento do risco de Sépsis/Meningite em prematuros de muito baixo peso. Por outro lado, as variáveis "$Peso$" e "$IdadeGestacional$" contribuem de forma negativa no aumento do risco de Sépsis/Meningite em prematuros de muito baixo peso.

Figura 7.1: Aplicação web do modelo de regressão logística múltiplo descrito na 5ª Etapa

Capítulo 11

Conclusões e desafios futuros

Ao longo deste trabalho foi notória a importância do *Jquery*, *Javascript*, *Html*, *Css* e *Bootstrap* cuja exploração desejo aprofundar num futuro próximo. Note-se, também, o quão útil se tornou o *latex* na escrita deste trabalho, o qual era desconhecido até então.

O contacto com o "mundo real" permitiu aumentar a importância de um trabalho autónomo mas também de partilha com todos os colegas, aperfeiçoar pesquisas de artigos e metodologias, aumentar ainda mais o grau de persistência na obtenção de melhores resultados, explorar ainda mais o R, melhorar a capacidade de adaptação a diferentes temáticas, relativizar alguns problemas que foram surgindo e aumentar o grau de foco ao que nos dedicamos.

Realço e agradeço, mais uma vez, o espírito de equipa que prolifera em toda a empresa Mhii e a motivação diária que me foi incutida.

Utilizei o *Shiny* para contruir a aplicação mas é meu objetivo explorar outros sistemas para desenvolvimento de aplicações *web*.

Desejaria completar a recolha de códigos *ICD9* já iniciada, junto da comunidade médica e construir e validar um modelo com melhor capacidade preditiva do risco de adquirir uma infeção em ambiente hospitalar, concluindo, assim, o trabalho a que me dediquei em grande parte do estágio curricular desenvolvido na Mhii.

Saliento a importância de solicitar junto da comunidade médica o máximo cuidado nos registos efetuados uma vez que a "qualidade" dos dados e a inexistência de incoerências têm um forte impacto nas previsões dadas pelos modelos construídos.

O modelo final obtido não é certamente um modelo excecional mas é um modelo do qual

muito me orgulho dado o relativo curto espaço de tempo em que foi desenvolvido e aprimorado assim como a forma como fui gerindo todos os contratempos que foram surgindo. É pois uma grande vitória, uma vez que temia-se não ser possível construir um modelo a partir da base de dados que foi objeto de estudo.

Após a devida análise do modelo final e consequente experimentação da aplicação web construída, verificou-se que a variável "$SRDiasVentilacao$" e os níveis das respetivas variáveis "$Sim-SurfactantePosterior$", "$Sim-ReanimacaoAposSalaPartos$", "$Sim-Ventilacao$" e parto não espontâneo contribuem para o aumento do risco de Sépsis/Meningite em prematuros de muito baixo peso. Por outro lado, as variáveis "$Peso$" e "$IdadeGestacional$" contribuem para a diminuição do risco de Sépsis/Meningite em prematuros de muito baixo peso.

É meu objetivo melhorar ainda mais o modelo que suporta a aplicação já apresentada, tendo em conta apenas os níveis das variáveis categóricas identificados com importância, juntamente com as devidas variáveis numéricas e analisar a reduzida influência dos 28 indivíduos referidos anteriormente (cujos *Dfbetas* estão acima do limite indicado) nas estimativas dos parâmetros do modelo. Para atingir este objetivo poder-se-iam utilizar métodos de seleção robustos e também técnicas robustas para o tratamento de valores omissos.

Como notas finais, reafirmo que reunir, organizar e interpretar dados para sustentar a tomada de decisões em instituições públicas e privadas está agora no topo dos meus objetivos profissionais.

Bibliografia

[1] Natália Cordeiro e Alexandre Magalhães. *Introdução à Estatística - Uma perspetiva química*. Ed. por Lidel. 2004.

[2] Bernard Flury. *A first course in Multivariate Statistics*. Ed. por Springer. 1997.

[3] Rui Campos Guimarães e José Sarsfield Cabral. *Estatística*. Ed. por McGraw-Hill. 2007.

[4] Joseph Hilbe. *Pratical guide to logistic regression*. Ed. por Taylor Francis. 2015.

[5] Arthur Hoerl e Robert Kennard. «Ridge Regression: applications to nonorthogonal problems». Em: *American Statistical Association* 12 (1970).

[6] Christoph Hornik, Trem Fort, Reese H. Clark, Kevin Watt, Daniel K. Benjamin Jr., P. Brian Smith e Michael Cohen-Wolkowiez. «Early and Late Onset Sepsis in Very-Low-Birth-Weight Infants from a Large Group of Neonatal Intensive Care Units». Em: *Early Hum Development* (mai. de 2012).

[7] David W. Hosmer, Stanley Lemeshow e Rodney X. Sturdivant. *Applied Logistic Regression*. Ed. por Wiley. 2013.

[8] João Marôco. *Análise Estatística com o SPSS Statistics*. Ed. por Report Number. 2014.

[9] Ana Mirco. *A Sépsis no Recém-Nascido*. Ed. por Centro de Informação do Medicamento. 2010.

[10] Kristen S. Montgomery. «Apgar Scores: Examining the Long-term Significance». Em: *The Journal of Perinatal Education* 9.3 (2000).

[11] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Ed. por Wiley. 1987.

[12] Thomas Ryan. *Modern Regression Methods*. Ed. por Wiley. 1997.

[13] George Seber e Alan Lee. *Linear Regression Analysis*. Ed. por Wiley. 2003.

[14] Robert Tibshirani. «Regression Shrinkage and selection via the Lasso». Em: *Journal of the Royal Statistical Society* 58 (1996).

[15] The Pennsylvania State University. *Lesson 5: Regression Shrinkage Methods*. `https://onlinecourses.science.psu.edu/stat857/node/137`. Accessed: 2017-06-05.

[16] Hui Zou e Trevor Hastie. «Regularization and variable selection via the Elastic Net». Em: *Journal of the Royal Statistical Society* 67 (2005).

Apêndice

Tabela A: Descrição das variáveis numéricas da base de dados em estudo

Nome da Variável	Breve descrição
MaeIdade	Idade da Mãe
IdadeGestacional	Idade gestacional em dias
NascimentoPeso	Peso do recém-nascido ao nascer em gramas
NascimentoComprimento	Comprimento do recém-nascido ao nascer em centímetros
NascimentoPerimetroCefalico	Perimetro cefálico do recém-nascido ao nascer em centímetros
Apgar1	Índice de Apgar no primeiro minuto de vida do recém-nascido
Apgar5	Índice de Apgar no quinto minuto de vida do recém-nascido
Apgar10	Índice de Apgar no décimo minuto de vida do recém-nascido
SRDiasVentilacao	Número de dias em Ventilação

Tabela B: Descrição das variáveis categóricas da base de dados em estudo

Nome da Variável	Descrição	Níveis da Variável e proporção
NascimentoOutborn	Se o recém-nascido nasceu no hospital hospital onde foi feito o registo	Inborn (95.1%) Outborn (4.9%)
Transporte	Se a mãe foi ou não transferida de outro hospital no período pré ou pós parto	Amissão Materna Direta (67.1%) Ex-Utero (4.9%) In-Utero (28%)
CuidadosPrenatais	Se a mãe recebeu ou não cuidados obstétricos na admissão para o parto	Desconhecido Não (3.7%) Não aplicável ()0.1% Sim (96.2%)
ConcepcaoAssistida	Se a conceção foi ou não medicamente assistida	Não (86%) Sim (14%)
CorticoidesPrenatais	Se houve ou não administração de corticóides antes do nascimento	Completo (66.3%) Desconhecido (1.2%) Não (9.4%) Parcial (23.1%)
PatologiasNaGravidez	Se a gravidez decorreu com alguma patologia materna	Sim (65.9%) Não (34.1%)
TipoDeParto	Tipo de parto	Cesariana (71.6%) Vaginal (28.4%)
MotivoDoParto	Motivo do parto	Espontâneo (47.3%) IVG (31.4%) Patologia Fetal (21.1%) Patologia Materna (0.2%)
Sexo	Sexo do recém-nascido	Feminino (47.7%) Masculino (52.3%)
Gemelar	Se o recém-nascido resulta ou não de gestação simples	Sim (35.4%) Não (64.6%)
ReanimacaoSalaPartos	Se ocorreu algum tipo de reanimação na sala de partos	Sim (64.1%) Não (35.9%)
ReanimacaoAposSalaPartos	Se ocorreu algum tipo de reanimação fora da sala de partos	Sim (86.7%) Não (13.3%)

Tabela B (cont): Descrição das variáveis categóricas da base de dados em estudo

Nome da Variável	Descrição	Níveis da Variável e proporção
MalformacaoCongenitaMajor	Se foi ou não diagnosticada ao recém-nascido alguma malformação congénita major	Sim (3.9%) Não (94.8%) Não aplicável (1.3%)
SurfactanteInicial	Se o recém-nascido recebeu ou não surfactante exógeno na sala de partos ou equivalente	Desconhecido (0.1%) Não (93.2%) Sim (6.7%)
SurfactantePosterior	Se o recém-nascido recebeu ou não surfactante exógeno durante o internamento	Sim (38.8%) Não (61.2%)
OxigenioDia28	Se o recém-nascido se encontrava ou não a receber suplemento de O2 no 28º dia de vida	Não (46.8%) Não aplicável (30.7%) Sim (22.4%)
CorticoidesDPC	Se foram administrados corticóides após o nascimento para tratar ou prevenir doença pulmonar crónica	Não (94%) Não aplicável (1.3%) Sim (4.7%)
DiagSdr	Presença ou não de Síndrome de dificuldade respiratória	Não (31%) Não aplicável (1.3%) Sim (67.7%)
DiagPneumotorax	Presença ou não de ar extrapleural diagnosticado por radiografia ou drenagem pleural	Não (93.7%) Não aplicável (1.3%) Sim (5%)
DiagPda	Se o recén-nascido teve ou não Persistência de ductos arteriosus (PDA) hemodinamicamente significativo	Desconhecido (1.1%) Não (74.2%) Não aplicável (1.4%) Sim (23.3%)
DiagNec	Se o recém-nascido cumpriu ou não a definição de Enterocolite Necrotizante (NEC)	Não (92.5%) Não aplicável (1.3%) Sim (6.2%)
DiagPerfuracaoGi	Se o recém-nascido teve ou não uma perfuração gastrointestinal focal isolada independe de NEC	Não (96.5%) Não aplicável (1.3%) Sim (2.2%)

Tabela B (cont): Descrição das variáveis categóricas da base de dados em estudo

Nome da Variável	Descrição	Níveis da Variável e proporção
PdaProfilatico	Se foi ou não administrada indometacina ou ibuprofeno após o nascimento para profilaxia de PDA	Não (98.5%) Não aplicável (1.3%) Sim (0.2%)
PdaTerapeutico	Se foi ou não administrada indometacina ou ibuprofeno após o nascimento para tratamento de PDA	Não (87%) Não aplicável (1.3%) Sim (11.7%)
CirurgiaPda	Se foi ou não realizada laqueação cirúrgica do canal arterial	Não (95.7%) Não aplicável (1.3%) Sim (3%)
CirurgiaNec	Se foi ou não realizada alguma intervenção para tratamento de enterocolite necrotizante (NEC)	Não (96.8%) Não aplicável (1.4%) Sim (1.8%)
CirurgiaMajorOutra	Se foi ou não realizada outra cirurgia major, para além das anteriores	Não (95.5%) Não aplicável (1.4%) Sim (3.1%)
Hpiv	Pior grau de hemorragia peri ou intraventricular (HIV)	0 (68.1%) 1 (12.1%) 2 (7.3%) 3 (7.4%) Não aplicável (0.1%) Desconhecido (4.9%)
Evhp	Se o recém-nascido teve ou não enfarte venoso hemorrágico periventricular (EVHP) associado à HIV	Desconhecido (0.5%) Não (21.5%) Não aplicável (73.4%) Sim (4.5%)
EvhpExtensao	Evhp: extensão	Bilateral com desvio (0.2%) Bilateral sem desvio (0.8%) Desconhecido (0.1%) Não aplicável (95.6%) Unilateral com desvio (2.4%) Unilateral sem desvio (0.9%)

Tabela B (cont): Descrição das variáveis categóricas da base de dados em estudo

Nome da Variável	Descrição	Níveis da Variável e proporção
DilatacaoVentricularPh	Se o recém-nascido teve ou não dilatação ventricular pós-hemorrágica	Desconhecido (0.3%) Não (21.7%) Não aplicável (73.5%) Sim (4.5%)
LpvGrau	Pior grau de leucomalácia periventricular (LPV) detetada	0 (80.6%) 1 (8.8%) 2 (1.9%) 3 (0.9%) 4 (0.5%) Não aplicável (5%) Sem informação (2.3%)
ExameOftalmologicoRopGrau	Pior grau de retinopatia da prematuridade (ROP)	0 (49.3%) 1 (9.8%) 2 (5%) 3 (2.8%) 4 (0.1%) 5 (0.1%) Não aplicável (32.9)
ExameOftalmologicoRopCirurgia	Se o recém-nascido foi submetido ou não a crio-cirurgia ou a tratamento laser para ROP	Não (5.3%) Não aplicável (92%) Sim (2.7%)
ExameOftalmologicoPlus	Se foi ou não diagnosticada ROP grau 2 ou 3 associadas a sinais de incompetência vascular	Não (6.2%) Não aplicável (92%) Sim (1.8%)
Apgar	Se o índice Apgar é ou não inferior a 6 aos 5 minutos de vida do recém-nascido	0-Não (94%) 1-Sim (6%)
SepsisMeningite	Diagnóstico compatível com sépsis ou meningite	0-Não (62.5%) 1-Sim (37.5%)
SepsisMeningitePrecoce	Diagnóstico compatível com sépsis precoce ou meningite	0-Não (89%) 1-Sim (11%)
SepsisMeningiteTardia	Diagnóstico compatível com sépsis tardia ou meningite	0-Não (69%) 1-Sim (31%)

Tabela B (cont): Descrição das variáveis categóricas da base de dados em estudo

Nome da Variável	Descrição	Níveis da Variável e proporção
RessuscitacaoOxigenio	Se o recém-nascido recebeu ou não algum suplemento de oxigenio na sala de partos	Desconhecido Não Sim
RessuscitacaoInsuflador	Se o recém-nascido recebeu ou não algum tipo de pressão positiva por máscara ou *prongs* nasais e insuflador manual, *neopuff* ou similar na sala de partos	Desconhecido Não Sim
RessuscitacaoEntubacaoEt	Se o recém-nascido recebeu ou não ventilação através de tubo endotraqueal na sala de partos	Desconhecido Não Sim
RessuscitacaoCompressaoCardiaca	Se foi ou não efetuada massagem cardíaca externa na sala de partos	Desconhecido Não Sim
RessuscitacaoAdrenalina	Se foi ou não ministrada adrenalina na sala de partos	Desconhecido Não Sim
SROxigenio	Se o recém-nascido recebeu ou não suplemento de oxigénio durante o internamento	Não Não aplicável Sim
SRCpap	Se o recém-nascido recebeu ou não CPAP nasal durante o internamento	Não Não aplicável Sim
SRVppni	Se o recém-nascido recebeu ou não algum tipo de ventilação por pressão positiva não invasiva, sem entubação endotraqueal	Não Não aplicável Sim
SRVentilacaoOxidoNitrico	Se o recém-nascido recebeu ou não algum tipo de ventilação com o uso de óxido nítrico	Não Não aplicável Sim
SRIppv	Se o recém-nascido recebeu ou não algum tipo de ventilação por pressão positiva, via tubo endotraqueal	Não Não aplicável Sim

Tabela B (cont): Descrição das variáveis categóricas da base de dados em estudo

Nome da Variável	Descrição	Níveis da Variável e proporção
SRVaf	Se o recém-nascido recebeu ou não ventilação de alta frequência, via tubo endotraqueal	Não Não aplicável Sim
SRVafni	Se o recém-nascido recebeu ou não ventilação de alta frequência não invasiva, via tubo endotraqueal	Não Não aplicável Sim